Structural determinants of mutability across cancer genomes

Submitted for the degree of Doctor of Philosophy

by

Ilias Georgakopoulos Soares

University of Cambridge

Wellcome Sanger Institute

Magdalene College

January 2019

Preface.

This thesis is submitted for the degree of Doctor of Philosophy at the University of Cambridge. I declare this is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated in the text. This document, in whole or in parts, has not been submitted for any other degree or diploma. It does not exceed the prescribed word limit for the Degree Committee for the Faculty of Biology.

Ilias Georgakopoulos-Soares

Cambridge, UK

January 2019

The thesis is dedicated to my parents, Giannis and Ana Maria! Their support and most importantly their love have always been beyond words to describe. Additionally, it is dedicated to my grandmother Maria Augusta Lopes and to the memory of my grandfather, Belchior Peixinho Soares.

Structural determinants of mutability across cancer genomes

Ilias Georgakopoulos-Soares

Summary

Cancer is a group of diseases which are characterised and actuated by somatic mutations. In cancer the distribution of mutations across the genome is inhomogeneous, with genomic and epigenomic features influencing mutational patterns. Previous studies have indicated that chromatin organization and replication time domains are correlated with and thus predictive of this variation. Here the role of alternative DNA structures was investigated across a multitude of whole-genome sequenced cancers.

Sequences that are predisposed to fold in alternative DNA structures can be identified by the primary DNA sequence of the human genome and are collectively known as non-B DNA motifs. More specifically, these include Z-DNA, G-quadruplexes, inverted repeats that can fold in cruciforms and hairpins, direct and short tandem repeats that can mediate the formation of slipped structures and a subset of mirror repeats that fold in intramolecular triple stranded DNA also known as H-DNA.

A systematic investigation of the association between each of those non-B DNA motifs and mutability was performed across thousands of whole genome sequenced tumours from different tissues. Non-B DNA motifs were more mutable than the surrounding regions and were found to be determinants of mutability across cancer types. Additionally, they could be used to predict variation in mutational density genome-wide. Exposed structural components and physical properties of non-B DNA motifs influenced the likelihood of mutagenesis, indicating that secondary structures are possibly causally implicated in mutagenesis. Furthermore, non-B DNA motifs increased the likelihood of recurrent mutations in the genome, which has direct implications for the identification of driver mutations in non-coding regions.

A detailed characterisation of indel mutagenesis was performed across the different cancer types. The analysis indicated the roles of different non-B DNA motif categories as well as sequence homologies in indel mutagenesis. In particular, sequence characteristics of a subset of non-B DNA motifs significantly influenced their relative mutational enrichment at specific indel categories. Finally, a method was developed to quantify replication and transcription strand asymmetries at indels systematically for the first time. As a result, mutational processes that are causally implicated in strand asymmetries at indels were identified and analysed. These included mismatch repair and transcription-coupled nucleotide excision repair both of which contributed to the observed transcriptional strand asymmetries for indels.

Acknowledgements.

The last four years have been an incredible journey. During these years, I acquired lifetime friends, expanded my knowledge and understanding and gained memories to cherish and reminisce forever.

First and foremost, I wish to thank my two supervisors, Serena Nik-Zainal and Martin Hemberg, for their advice, insights and immense patience with me. They were not only supervisors, they were great role models as well as amazing scientists. I was extremely fortunate to have them and wish after the end of this journey to remain in close contact with both of them and continue discussing with enthusiasm and limitless energy scientific concepts and ideas. In addition, I would like to thank everyone in the labs of Serena Nik-Zainal and Martin Hemberg for their constructive feedback during presentations and fruitful discussions.

I would further like to thank my close friend and collaborator Naman Jain. He has been a great influence and helped me acquire programming and computational skills that proved vital in the later steps of my graduate studies. In addition, I would like to thank Sandro Morganella for his help in mutational analyses and for kindly sharing code and data with me, which proved crucial for my PhD project progression, and which were given to me without any hesitation.

A big thank to all my new friends that I have acquired in Magdalene college, in the Wellcome Sanger Institute and others I met during my studies in Cambridge University including my great friend Evangelos Avgoulas, also known as the magician who has introduced me to the world of magic and illusion, my favourite mexican friend known as Guillermo Parada, my neighbour Nikita Makarchev, my great friend Aikaterini Chatzipli, Britney Marshall also known as the bee, my gym buddy Grigoris Gakis, Ami Day, Allan Muhwezi, Yuan He and Reelika Priimägi to name only a few. I would also like to thank my close friend Tasos Tsegas who has always been my great friend from childhood to present and my close friend Apostolos Patsakis, who is becoming stronger every day.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

1. Introduction

"*If you want to understand function, study structure.*"
Francis Crick

1.1. Somatic mutations in cancer genomes.

The human body is composed of 37 trillion cells (Bianconi et al. 2013). Throughout life, DNA in each cell of the human body is subjected to a wide variety of DNA damaging events. As an average, each cell receives 19,000 DNA damage incidences per day (Vilenchik et al. 2000), although this number is highly variable depending on the tissue type. For instance, skin cells are exposed to an excess of exogenous DNA damage primarily induced by UV light, while reactive oxygen species are an endogenous source of genomic DNA damage. As a result, human cells require a number of repair pathways to repair DNA damage continuously and to maintain genome integrity. Damage that is not corrected, however, will become fixed as somatic mutations, which will accumulate throughout the lifetime of a person.

Cancer is a disease that is characterized and actuated by somatic mutations (Stratton et al. 2009). The tumour mass originates from a single common ancestor cell and during its growth it continuously acquires new mutations. Therefore, the cancer genome differs from the genome of the other cells in a human body because it contains the set of mutations that have accumulated in the process of cancer development and progression. The accelerated progress in sequencing technologies has potentiated the sequencing of whole genomes with diminishing costs. The analysis of originally few cancer genomes (Pleasance et al. 2010a), (Pleasance et al. 2010b), (Nik-Zainal et al. 2012a), (Nik-Zainal

et al. 2012b) and lately hundreds or thousands of cancer genomes across multiple tissues (Nik-Zainal et al. 2016), (ICGC and TCGA projects), (Campbell et al. 2017) has advanced our understanding of mutagenesis during cancer development, and has elucidated the mechanisms of a number of DNA damage and repair processes.

1.1.1. DNA damage and repair in mutagenesis.

DNA damage can originate by a number of different processes, many of which are described below. Depending on the type of DNA damage, different DNA repair pathways are activated to ensure the integrity of the genome. However, if the DNA damage cannot be repaired or the mutational burden is high, cells can undergo apoptosis resulting in the death of the cell or cellular senescence at which stage, the cell is unable to further proliferate (d'Adda di Fagagna and di Fagagna 2008).

Spontaneous and enzymatic DNA damage. Spontaneous damage results in abasic sites which depending on the damaged nucleotide can be further classified into apurinic and apyrimidinic sites. Such lesions are resolved by base excision repair (BER) (Wallace 2014). The damaged nucleotide is repaired using the corresponding base of the opposite strand to identify the missing nucleotide. Deamination of bases by enzymatic processes also results in DNA damage. For instance, the family of APOBEC deaminases primarily deaminate viral material therefore serving as protective agents in human cells; however, they can also be described as a double-edged sword since they can cause collateral DNA damage in the human genome (Roberts et al. 2013). Single stranded DNA in the genome, which can form during transcription and replication among other processes, is a preferable target of APOBEC deamination. In particular, *kataegis* mutagenesis, which involves a pattern of localised hypermutation, is hypothesized to be driven by APOBEC mutagenesis (Nik-Zainal et al. 2012a).

Radiation. Radiation can be separated into ionizing and non-ionizing depending on the length of the wavelength. Ionizing radiation has short wavelength and high energy, and causes frequent double strand breaks (DSB), which if unresolved can result in gross indels and rearrangements. The two main repair pathways that fix double-strand DNA breaks include: i) Homologous recombination (HR) and ii) Non-homologous end joining (NHEJ) (Jasin and Rothstein 2013), (Prakash et al. 2015), (Chang et al. 2017). Both mechanisms use microhomology patterns at the site of double-strand break. HR is usually error-free and uses longer homology patterns to an undamaged DNA template, whereas NHEJ is error-prone and uses shorter microhomology patterns at the site of the double-strand break. Non-ionizing radiation on the other hand has longer wavelength and lower energy, and causes single or double nucleotide mutations (Pfeifer et al. 2005). One common source of non-ionizing radiation is UV-light, which causes an excess of mutations in the genome of cells found in exposed areas of the human body, such as the skin.

Reactive oxygen species and other chemical agents. Endogenous reactive oxygen species (ROS) can interact with the DNA molecule resulting in single-strand breaks (SSB) and abasic sites. Multiple repair pathways are implicated in the repair of such damage events including BER and nucleotide excision repair (NER), (Wallace 2014), (Marteijn et al. 2014). Numerous different ROS have been described and the type of DNA damage and repair is dependent on the particular chemical and its properties. An excess of ROS is produced in mitochondria during physiological conditions and they have been implicated in aging (Bratic and Larsson 2013). Other chemical agents that damage the DNA include alkylating agents, platinum-based compounds, poly-aromatic hydrocarbons, intercalating agents and psoralens among many. For instance, polyaromatic hydrocarbons include tobacco-smoke related carcinogens that result in a characteristic mutational pattern in lung cancers (Alexandrov et al. 2016).

Exogenous chemical sources of mutagenesis through adduct formation. Tobacco-smoke related carcinogens (e.g. polycyclic aromatic hydrocarbons) and aristolochic acid are exogenous chemical agents of mutagenesis that mediate their mutational effects through

the formation of DNA adducts. Aristolochic acid is found in the plant *Aristolochia* and has been used in traditional herbal medicines. It is a potent mutagen and a characteristic combination of mutation patterns known as a mutational signature has been observed and characterized, following exposure to it (Hoang et al. 2013), (Poon et al. 2015). Similarly, a mutational signature has been associated to tobacco-smoke in lung cancers (Alexandrov et al. 2013). Nucleotide excision repair is a group of repair pathways commonly separated in global genome nucleotide excision repair (GG-NER) and transcription-coupled nucleotide excision repair (TC-NER). TC-NER is a repair pathway that preferentially repairs DNA damage, such as bulky adducts, in the template strand. As a result, an enrichment of substitutions related to DNA damage through bulky adduct formation is observed in lung cancer genomes in the non-template strand at transcribed regions (Pleasance et al. 2010b). A transcriptional strand asymmetry is also observed in aristolochic acid carcinogenesis, suggesting that TC-NER is involved in repair of aristolochic acid damage preferentially in the template strand. As a result, DNA damage and repair processes leave their characteristic imprints in the genome, which can be identified.

1.1.2. Patterns of mutagenesis in the genome.

Somatic mutations are not randomly distributed in the genome with the sequence context, DNA accessibility and epigenomic landscape being major contributors of the likelihood of mutagenesis. Patterns of mutations represent traces of DNA damage and repair processes that have been operative, with each mutational process generating a characteristic and distinct pattern of mutations; its mutational signature. As a result, the cancer genome can be described as a valuable archaeological record from which we can gain insight into the mechanisms of mutagenesis. With recent algorithm developments, a multitude of mutational signatures have been extracted, many of which have been described in detail while others remain of unknown origin (Alexandrov et al. 2013), (Nik-Zainal et al. 2016), (Alexandrov et al. 2018).

Also, rearrangement signatures have been extracted; initially across 560 whole genome sequenced breast cancers and more recently across a plethora of cancer types. Therefore, it was demonstrated that imprints of mutational processes in the cancer genome can be identified at the rearrangement level (Nik-Zainal et al. 2016).

Additionally, classification of deletions based on the mechanism of their formation into repeat-mediated and microhomology-mediated has been used to gain insights into mutational processes (Nik-Zainal et al. 2016). Repeat-mediated deletions are flanked by a repeat tract with which they share the same repeating unit, are usually small (<3bp) and arise from erroneous repair of insertion-deletion loops. In contrast, microhomology-mediated deletions show homology of several nucleotides between deletion and the flanking sequence and are usually larger (>=3bp). Microhomology-mediated deletions are enriched in *BRCA1* and *BRCA2* deficient cancer patients which lack homologous recombination (HR). In the absence of HR, cells resort to alternative repair pathways such as non-homologous end joining (NHEJ), which tends to be error prone. On the other hand, repeat-mediated deletions are present in excess in mismatch repair deficient samples and usually occur at repeat sequences.

A number of clinical applications based on mutational signatures are being advanced. Firstly, mutational signatures could serve as a highly accurate diagnostic tool. Such an example has been the development of HRDetect. This machine-learning algorithm utilises mutational signatures to detect *BRCA1* and *BRCA2* deficient patients (Davies et al. 2017). Secondly, they can be utilised to advance group or personalised treatments. For instance, mismatch repair deficiency, which is identifiable through mutational signatures across different tissues, can identify highly responsive patients to immunotherapy (Le et al. 2017). Finally, mutational signatures could be used to identify currently unknown mutagenic sources and to increase our understanding of lifestyle factors that can be modified for cancer prevention. Therefore, mutational signatures could provide insight into yet unknown mutagenic factors and could serve as actionable targets for personalised therapies (Nik-Zainal and Morganella 2017).

1.1.3. Passenger and driver mutations in cancer.

Most mutations do not confer a selective advantage and are therefore termed passenger mutations. Driver mutations are the subset of mutations that confer selective advantage during cancer progression (Stratton et al. 2009). Recurrent mutagenesis analyses have been implemented to identify potential driver mutations.

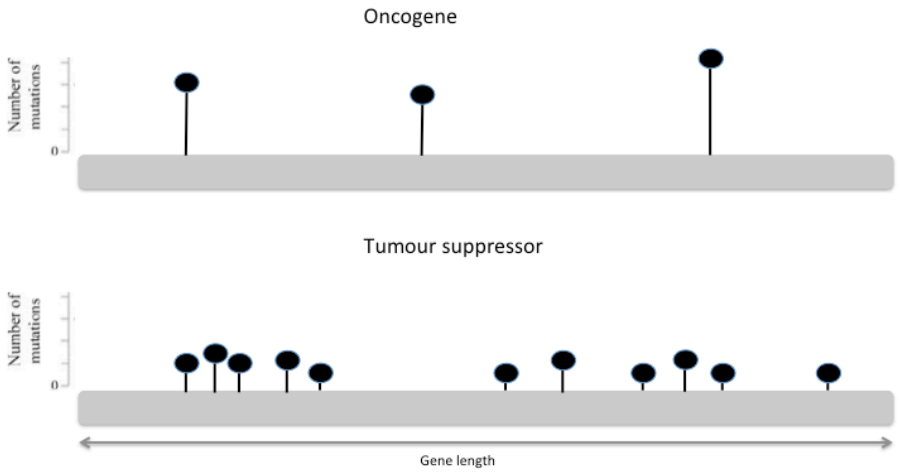
For the coding part of the genome, we understand the relationship between trinucleotides or codons at the DNA level and amino acids at the protein level. This helps explain the causal relationship between a recurrent coding mutation and its predicted effect at the protein level. Driver mutations disrupt normal protein expression or structure. Only a handful of driver mutations are usually required for cancer development (Stratton et al. 2009).

Several methods have been developed to distinguish and categorise drivers of cancer development. Two categories of cancer genes include oncogenes and tumour suppressors, at which driver mutations often result in activating or inactivating mutations respectively. Oncogenic driver mutations usually tend to be limited to the same amino acid positions, whereas tumour suppressor driver mutations are dispersed across the gene length (Figure 1.1a) and are often non-sense or frame-shifting, disruptive mutations (Vogelstein et al. 2013).

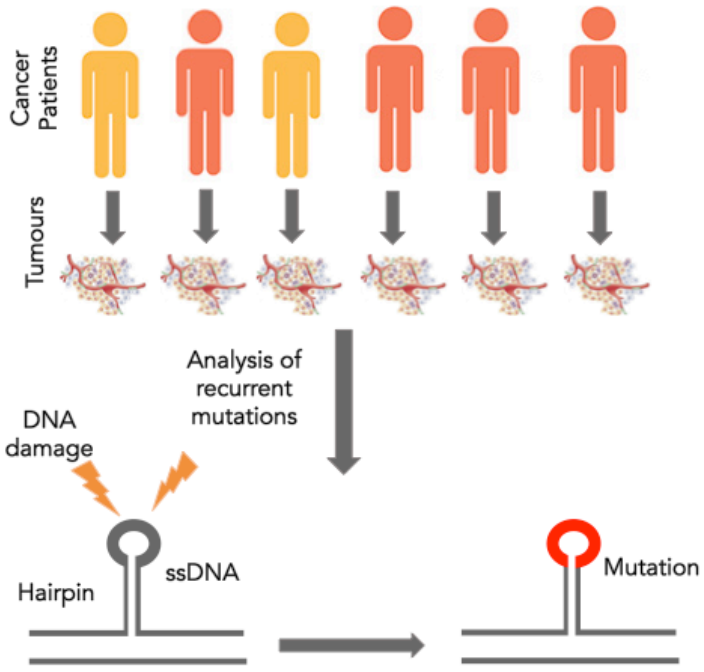
Driver genes are detected because they are mutated in excess across cancer patients. For instance, mutations of the tumour suppressor *TP53* are found in approximately 50% of cancer patients (Vogelstein et al. 2000) and can be identified by simple statistical analysis. Nevertheless, other putative driver genes are mutated in a smaller subset of patients and require more complex models of background mutability rates and larger patient cohorts to be identified (Wood et al. 2007). Models that quantify evolutionary

selection in cancer development have also been implemented in the search for driver mutations (Greenman et al. 2006). As a result, sites of recurrent mutagenesis that result from increased likelihood of mutagenesis can be distinguished from those that result through selection and are driver mutations. For instance, methods that use the ratio of non-synonymous to synonymous mutations (dN / dS) and incorporate the trinucleotide context in the model successfully identify genes under selection (Martincorena et al. 2017). So far numerous whole-exome sequencing studies have focused on characterising recurrent mutations in coding sequences to identify and explore putative driver mutations.

However, the rules that govern the functional roles of non-coding regions remain largely unknown. There is a rapidly growing number of whole genomes available allowing for the analysis of recurrent mutagenesis across the genome in patient cohorts. Recurrent mutations in the non-coding part of the genome, could be driver mutations that confer selective advantage or passenger mutations which occur at hypermutable loci. For instance, driver mutations at the *TERT* promoter, which encodes for telomerase reverse transcriptase, have been described across different types of tumours (Weinhold et al. 2014). However, finding and interpreting driver mutations in non-coding regions has remained more challenging. Currently differences in the likelihood of mutagenesis across the human genome are often ignored. However, in human cancers, recurrent mutagenesis reveals hints of relationships between secondary structure formation and an increased mutability in somatic cells, as most notably described for *PLEKHS1* promoter (Weinhold et al. 2014), (Nik-Zainal et al. 2016), (Figure 1.1b). Therefore, it could be the case that structural properties at the DNA level predispose it to the likelihood of mutagenesis.



a.



b.

Figure 1.1: Recurrent mutagenesis in cancer.

a. Schematic representation of driver mutations in oncogenes and tumour suppressors. Driver mutations in oncogenes are found recurrently at specific positions across the gene length, whereas in tumour suppressors they are usually dispersed across the gene length. b. Secondary structure formation results in increased propensity of DNA damage and recurrent mutagenesis across multiple cancer patients.

1.2. Secondary structures in the genome.

In a landmark paper that shaped biology, Watson and Crick described the double helical structure of the DNA molecule (Watson and Crick 1953). Double stranded DNA (dsDNA) is composed of two helices held together by hydrogen bonds and most often adopts the canonical right-handed double helical secondary structure, also known as the B DNA form. Nevertheless, it was later realised that alternative conformations of the DNA could be found in cells. A number of different biological processes, such as transcription, replication, recombination, DNA damage and repair can change the structure of the DNA, either transiently or for longer periods. These processes distort the B DNA form of the DNA molecule, and are favourable for alternative DNA conformations, collectively termed non-B DNA structures.

A multitude of studies illustrated the formation of numerous non-canonical DNA structures across diverse organisms and conditions at the DNA level. Currently, there are more than 20 non-canonical secondary structures known that can form at DNA (Ghosh and Bansal 2003). DNA supercoiling measures twisting against the helical conformation, with positive and negative supercoiling referring to overwinding and underwinding respectively. The conditions necessary for the formation of these structures vary; however negative supercoiling of the DNA has a pivotal role and can induce the formation of the majority of these structures (Sinden and Pettijohn 1984), (Herbert and Rich 1996), (Sun and Hurley 2009), (Brooks and Hurley 2010).

In this work, I focus on non-B DNA motif sequences which have been shown to form secondary structures and whose positions in the genome can be predicted from the primary nucleotide sequence of the human reference genome. These DNA motif sequences are: G-quadruplexes, inverted repeats, mirror repeats, H-DNA, direct repeats, short tandem repeats and Z-DNA, each of which is described below in greater detail.

1.2.1. Types of secondary structures and physical characteristics.

G-quadruplex.

G-quadruplex is the most thoroughly studied non-B DNA motif. It is found in GC-rich regions of the genome and is held together by Hoogsteen hydrogen bonds, which are a form of non-Watson-Crick hydrogen bond base pairing (Bochman et al. 2012), (Chen and Yang 2012), (Kwok and Merrick 2017). These bonds connect four guanines forming a square planar arrangement (G-quartet) (Figure 1.2b). Stacking of multiple G-quartets results in the formation of the G-quadruplex structure, while certain monovalent cations favour the G-quadruplex structure stabilization (Sen and Gilbert 1990). The intervening sequences between the G-quartets form single stranded loops. G-quadruplex structures can be generated by a single DNA strand (intramolecular G-quadruplex) or by multiple DNA strands (intermolecular G-quadruplex), and are termed unimolecular G-quadruplexes when they are generated by a single strand, bi-molecular when two strands are used and tetramolecular when four strands are contributing to their formation (Burge et al. 2006). They can also be classified into parallel, antiparallel or hybrid G-quadruplexes depending on the folding topology (Burge et al. 2006), (Figure 1.2c). In addition, recent work has shown the existence of more complex arrangements (Lim et al. 2015), (Bartas et al. 2018).

Specific properties such as length of motif, length of G-runs and loops as well as nucleotide composition determine G-quadruplex stability and the likelihood of formation (Tippana et al. 2014), (Piazza et al. 2015), (Kim et al. 2016). Smaller loops result in higher stability of G-quadruplexes (Hazel et al. 2004), (Rachwal et al. 2007), (Bugaut and Balasubramanian 2008), (Huppert 2010), (Guédin et al. 2010). Similarly, longer G-runs allow for more stable G-quadruplex formation (Huppert 2010). The nucleotide composition of the looping region also determines their folding potential and stability (Tippana et al. 2014), (Piazza et al. 2015), (Kim et al. 2016). Therefore, the primary

nucleotide sequence of the G-quadruplex dictates its folding kinetics and thermodynamic stability.

Intramolecular G-quadruplexes are the most well described type of G-quadruplex. It has been noted that a sequence motif can capture the majority of the sites in the genome that can form intramolecular G-quadruplexes accurately (Huppert and Balasubramanian 2005). The consensus G-quadruplex motif: $G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}$ has been used to quantify the number and distribution of G-quadruplexes in the genome and is utilised in numerous available methods (Kikin et al. 2006), (Huppert and Balasubramanian 2007), (Zhang et al. 2008), (Wong et al. 2010), (Cer et al. 2011), (Cer et al. 2013), (Figure 1.2a). Nevertheless, it was quickly found that not all intramolecular G-quadruplexes follow the same rules and exceptions have been pointed out that do not adhere to the consensus motif (Phan et al. 2007), (Mukundan and Phan 2013), (Jodoin et al. 2014), (Martadinata and Phan 2014), (Onel et al. 2016), (Varizhuk et al. 2017). These include G-quadruplexes that contain G-runs with interruptions (bulges), mismatches or longer loops (Amrane et al. 2012), (Agrawal et al. 2014), (Jodoin et al. 2014). As a result, the frequency and diversity of G-quadruplexes in the genome could be currently substantially underestimated.

Intermolecular G-quadruplexes have also been shown to form at the DNA level, the RNA level or a hybrid involving both DNA and RNA (Wanrooj et al. 2012), (Zheng et al. 2013). Nascent RNA transcripts can facilitate the formation of intramolecular and intermolecular G-quadruplexes with catalytic roles *in vitro* (Shrestha et al. 2014). Importantly, locations in the human genome where intermolecular G-quadruplexes could form are extremely prevalent (97% of human gene promoters) and show strand asymmetry in their distribution relative to the transcription start site (TSS) (Zheng et al. 2013). DNA-RNA interactions are stronger and more stable than DNA-DNA interactions, which could imply that once these structures form at promoters they can steadily regulate gene expression levels. Furthermore, RNA-DNA hybrid G-quadruplex formation has been shown to be induced by transcription (Shrestha et al. 2014).

Evidence for G-quadruplex formation. Observing G-quadruplex formation *in vitro* and in *vivo* has been challenging. Initial observations of G-quadruplex formation were reported in 1988 (Sen and Gilbert 1988). A number of different methods have been used to study the G-quadruplex structure. These include X-ray crystallography, circular dichroism and NMR spectroscopy among others (Dai et al. 2006), (Chen and Yang 2012), (Bochman et al. 2012). These methods provided insight into the three-dimensional G-quadruplex structures, including the different conformations such as parallel and antiparallel orientation and the role of the loop size in its formation (Wang and Patel 1993), (Phan et al. 2007), (Campbell and Parkinson 2007), (Vorlíčková et al. 2012). Additionally, UV-spectroscopy and FRET melting have been applied to study the thermostability of G-quadruplexes (Mergny et al. 1998), (Mergny et al. 2001). Nevertheless, additional evidence was required to convince sceptics regarding G-quadruplex formation *in vitro* and in *vivo*.

Techniques used to visualise G-quadruplexes in cells have provided the strongest evidence for their formation *in vivo*. Antibodies that are specific to G-quadruplex have been raised and can be used to pull down G-quadruplex DNA (Lam et al. 2013) or implemented with fluorophores (Biffi et al. 2013). Fluorescence microscopy has been employed to detect their presence and localisation inside cells. By implementing these advances, ChIP-seq experiments targeting G-quadruplexes have generated genome-wide maps of their distribution, while a G-quadruplex sequencing method has been reported. This method uses G-quadruplex-dependent polymerase stalling followed by high throughput sequencing, to generate genome-wide maps of the occurrences of G-quadruplex structures in the genome (Chambers et al. 2015). Results from G4-seq have been extremely interesting; besides identifying an excess of 700,000 G-quadruplex positions in the genome in which G-quadruplex structures could potentially form, they also found a proportion of those sites with the potential to form G-quadruplexes that did not match the consensus G-quadruplex motif. In addition, recently G-quadruplexes have been observed in human cells (Biffi et al. 2013), (Henderson et al. 2014) and *in vivo* in zebrafish (Agarwal et al. 2014) providing further support for their formation and functionality.

Stabilisation of G-quadruplexes. Multiple compounds have been shown to stabilise G-quadruplex structures. These include monovalent cations (K^+>Na^+>NH_4^+>Li^+) (Davis 2004) and G-quadruplex ligands. A commonly used ligand for G-quadruplex stabilisation has been pyridostatin (PDS), which increases their formation by 4.8-fold (Rodriguez et al. 2008). A derivative of PDS, named carboxypyridostatin (cPDS) binds preferentially to G-quadruplexes generated in the RNA level (Di Antonio et al. 2012), and allows for the investigation and discrimination of the G-quadruplex roles in the DNA and RNA level separately (Rocca et al. 2017).

i-motifs. i-motifs can sometimes form in the complementary single stranded part of the G-quadruplex, which is C-rich. They are formed by cytosine-cytosine base pairing held in a quadruplex formation (Gehring et al. 1993). Properties such as loop length, sequence composition and environmental pH affect their formation dynamics (Lieblein et al. 2013), (Gurung et al. 2015), (Takahashi and Sugimoto 2015), (Reily et al. 2015), (Dzatko et al. 2018). In particular, i-motifs are thought to be more unstable than G-quadruplexes, because they require acidic pH for their formation. Because i-motifs are stabilised by acidic pH they have been used in nanotechnology applications as switches based on that particular property (Dong et al. 2014).

Recently a study provided *in vivo* evidence for i-motif formation in human cells, especially at regulatory regions and telomeres, using an antibody fragment that directly targets them (Zeraati et al. 2018). Nevertheless, i-motifs are more unstable than G-quadruplexes and remain an unexplored secondary structure with a plethora of potential regulatory roles. Since G-quadruplexes have been shown to contribute in numerous biological processes already and since both G-quadruplexes and i-motifs can form at the same positions in the genome, it is plausible that i-motifs have yet unknown functions at the same or different processes.

R-loops. R-loops consist of a DNA-RNA hybrid, in which the RNA is held together with one DNA strand, while the other DNA strand remains single stranded (Santos-Pereira

and Aguilera 2015). Overall, R-loops are GC-rich. GC-skew levels can be used to find regions with higher likelihood of R-loop formation, but the mapping ability has remained poor. R-loops are enriched at G-quadruplex sites, which also show strong GC-skew and they can stall the RNA polymerase progression. R-loops have received a lot of recent attention, partly because of progress in the field to identify them more effectively and partly because of discoveries relating them to mutagenesis and regulation of gene expression (Chen et al. 2017), (Dumelie and Jaffrey 2017). In particular, novel methods such as DRIP-seq and bisDRIP-seq have been developed to map R-loops with higher precision genome-wide (Dumelie and Jaffrey, 2017). As their roles in the genome are starting to be unraveled, it is found that promoter regions are enriched in R-loops in which they have regulatory functions (Ginno et al. 2012). Additional roles could involve genome integrity; for instance, the severity of the effect of head on replication-transcription collisions is influenced by the presence of R-loops (Hamperl et al. 2017).

*Functional roles of G-quadruplexes***.**

Gene regulation. About half of human genes have a G-quadruplex within 1,000nt upstream of the TSS (Huppert and Balasubramanian 2007). At promoters, G-quadruplexes are more frequently found in nucleosome-free regions and could therefore be involved in transcription factor recruitment (Hershman et al. 2008). Their transcriptional roles can be different and even opposing depending on their orientation relative to transcription and to their position within the gene (Brooks and Hurley 2010), (Lam et al. 2013), (Armas et al. 2017), (Bay et al. 2017). Downstream of the TSS, certain roles in transcription regulation mediated by G-quadruplexes involve: i) interference with the RNA polymerase (polymerase stalling) if they are formed in the template strand (Armas et al. 2017), ii) allowing the DNA to remain open for longer if they are found in the non-template strand (Bochman et al. 2012), iii) interactions between them and proteins. G-quadruplexes in the 5'UTR have also roles in the repression of translation (Halder et al. 2009), (Beaudoin and Perreault 2010). (Figure 1.2d). Additionally, G-quadruplexes have roles in transcription termination, polyadenylation and mRNA stabilisation (Beaudoin et al. 2013). Interestingly, a single G-quadruplex is a major regulator of polycistronic

transcription and replication in human mitochondria (Wanrooij et al. 2010), (Wanrooij et al. 2012).

Splicing. Splicing is a major post-transcriptional regulatory process which regulates the formation of the mature mRNA transcript that is used for translation. G-quadruplexes have been found in splicing sites in a number of genes. G-quadruplexes act as intronic splicing enhancers in the case of *hTERT* (Gomez et al. 2004) and their presence is also linked to exon exclusion, as in the case of *TP53* (Marcel et al. 2011). Furthermore, masking splicing signals by the formation of various secondary structures including G-quadruplex, could be another mechanism by which they act to determine splicing events (Chen and Mandley 2009).

Telomeres. Telomeres are heavily studied and their length is correlated to the number of cell cycles a cell has experienced. In multiple cancer types, telomere length is extended, allowing further proliferation of cancer cells. This feature contrasts the shortening of telomeres after each cell division of somatic cells, eventually resulting in their senescence or death. G-quadruplexes are involved in the formation of the telomere-dependent bouquet structure (Sen and Gilbert 1988). In telomeres, intramolecular G-quadruplexes have an anti-parallel form and block the polymerase activity. In contrast to that, intermolecular G-quadruplexes are of parallel form and allow for polymerase activity (Zahler et al. 1991), (Oganesian et al. 2006), (Oganesian and Bryan 2007). The role of G-quadruplexes in modulating telomerase activity has received considerable attention and has been thought of as a target for anticancer therapies (Neidle 2010). This suggests that non-B DNA conformations in the genome are druggable and perhaps actionable targets, although further investigation would be required.

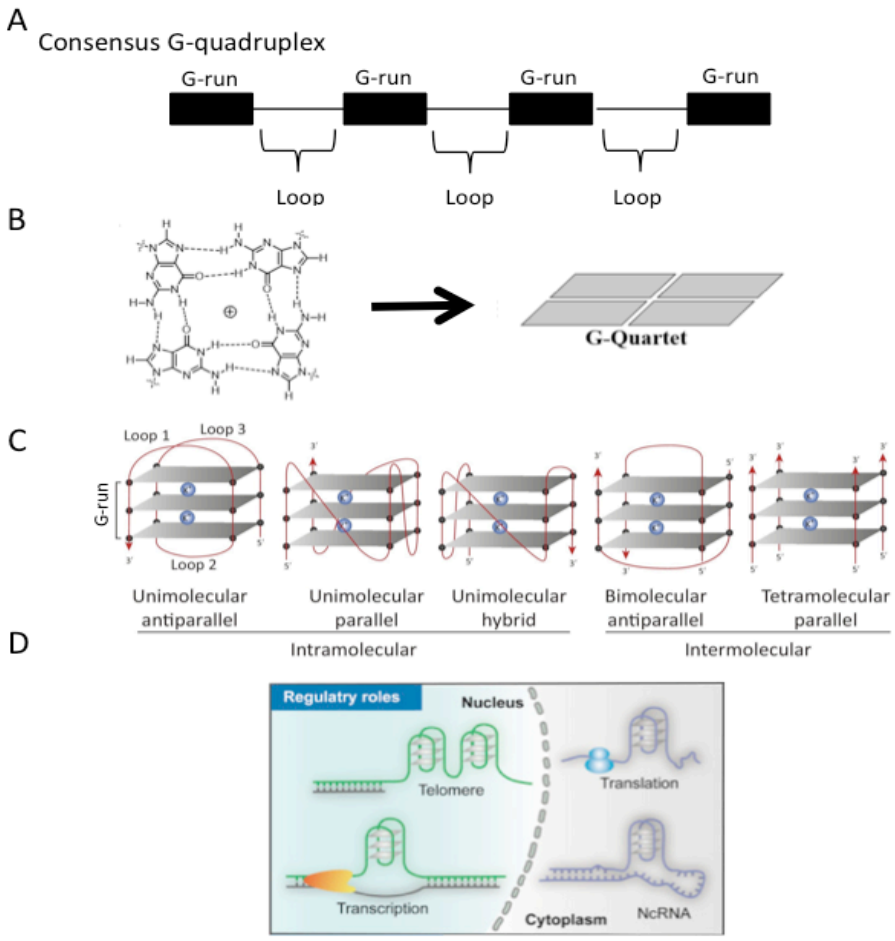


Figure 1.2: Schematic of G-quadruplex structure and functions.

a. Consensus motif for the identification of putative G-quadruplex structures. b. G-quartet stabilisation by Hoogsteen bonds. Schematic adjusted from (Che et al. 2018). c. Representative topologies of G-quadruplex structure. Schematic from (Kwok and Merrick 2017). d. Regulatory roles of G-quadruplexes in telomeres, transcription regulation, translation and non-coding RNAs. Schematic from (Tian et al. 2018).

Inverted repeats.

Inverted repeats are composed of two adjacent copies of the same sequence, one of which is found in the reverse complement orientation. The two copies are termed arms. An intervening non-complementary sequence termed spacer can separate the two arms (Figure 1.3a). If there is no spacer sequence between the two arms, the inverted repeat

is often called a palindrome. Inverted repeats can be further divided into perfect and imperfect. Perfect inverted repeats do not allow for mismatches in the arms, in contrast to imperfect inverted repeats, which can contain disruptions and mismatches.

Inverted repeats can fold primarily into two secondary structures; hairpins and cruciform (Kurahashi et al. 2004), (Mikheikin et al. 2006), (Zhao et al. 2010). A hairpin is held together by hydrogen bonds between the two complementary arms, while the spacer remains single stranded. Cruciforms consist of two hairpins and a 4-way junction, and resemble the Holliday junction which forms during recombination (Watson et al. 2004). In both hairpins and cruciforms, the spacer sequence remains single stranded and exposed whereas the arms base pair with hydrogen bonds and remain double stranded.

It has been shown that specific properties of inverted repeats, including spacer and arm length, interruptions and nucleotide composition can all affect the likelihood of secondary structure formation, its folding kinetics and its stability. Inverted repeats with an arm length of seven or more nucleotides have been shown to form *in vivo* (Nag and Petes 1991). Hairpin formation dynamics have been studied in detail by varying the spacer and arm lengths and their nucleotide composition and examining the folding and mutagenic potential (Varani 1995), (Nag et al. 1997), (Lobachev et al. 1998), (Woodside et al. 2006). For instance, arms with higher GC content display more stable formation (Woodside et al. 2006). Inverted repeats with imperfections have higher energy of cruciform extrusion (Benham et al. 2002). Additionally, inverted repeats with no spacer sequence, also known as palindromes, are more likely to fold in cruciform structures, whereas inverted repeats with a spacer sequence of at least four nucleotides favour the formation of hairpins (Nag and Petes 1991), (Varani 1995).

Shorter inverted repeats are more stable than the longer counterparts, which are more prone to mutagenesis (Sinden et al. 1991). Due to their inherent instability, long, perfect inverted repeats are rare in the human genome (Lobachev et al. 2000). Attempts to insert long inverted repeats in plasmids have led to their elimination within few generations due to their inherent instability (Collins 1981), (Leach 1994), (Lobachev et al. 1998),

(Kurahashi 2001). In contrast, short inverted repeats are prevalent in prokaryotic and eukaryotic genomes (Cox and Mirkin 1997), (Lillo et al. 2002), (Ladoukakis and Eyre-Walker 2008), (Strawbridge et al. 2010). In particular, they are much more prevalent than would be expected by chance in a number of different genomes including both prokaryotic and eukaryotic organisms (Cox and Mirkin 1997), (Lilo et al. 2002), (van Noort et al. 2003), (Ladoukakis and Eyre-Walker 2008).

Evidence for secondary structure formation at inverted repeats. A number of different methods have been applied to investigate secondary structure formation at inverted repeats. Cruciform structures were initially identified by S1 nuclease probing (Lilley 1980), (Panayotatos and Wells 1981). Since then, cruciform structures have been studied systematically with gel electrophoresis *in vitro* (Lyamichev et al. 1983). Psoralen is a molecule that can intercalate at DNA regions with negative supercoiling and with UV-light exposure can crosslink complementary DNA strands. Cruciform structures have been detected to form *in vivo* using psoralen in psoralen crosslinking assays (Zheng and Sinden 1988), (Zheng et al. 1991), (Sinden et al. 1991). More recently, psoralen assays have been used to investigate genome-wide differences of DNA supercoiling *in vivo* (Naughton et al. 2013), (Kouzine et al. 2013).

Antibodies have been generated against cruciforms for their identification (Zannis-Hadjopoulos et al. 1988). Microscopy techniques such as Atomic Force Microscopy (AFM) have been used to visualise cruciforms. Electron microscopy has also been used for the identification of hairpin and cruciform formation at inverted repeats (Kato et al. 2003), (Kurahashi et al. 2004), (Mikheikin et al. 2006). A multitude of studies have also stretched the role of supercoiled DNA on cruciform formation (Liley 1980), (Panayotatos and Wells 1981), (Mizuuchi et al. 1982). As a result, a plethora of evidence has accumulated for the potential of inverted repeats to fold in secondary structure formations.

Recent studies have analysed G-quadruplex sequences that contain an inverted repeat within a long loop (>=7nt) of the G-quadruplex structure (Lim and Phan 2013), (Benabou et al. 2014), (Lim et al. 2015). These structures are found in regulatory regions and are

very stable. The inverted repeat within it plays a pivotal role during folding and formation of the structure, therefore implicating combinations of secondary structures cooperatively (Risitano and Fox 2003), (Lim et al. 2013).

Multiple programs and databases have been developed to identify inverted repeats in genomic sequences (Rice et al. 2000), (Warburton et al. 2004), (Cer et al. 2013), (Ye et al. 2014), (Brázda et al. 2016), (Baskett et al. 2017), (Wang and Huang 2017). These programs differ in the type of secondary structure they focus on based on the primary nucleotide sequence, the size of inverted repeats that can be identified and some allow for search of imperfect inverted repeats and long inverted repeats with multiple disruptions.

Functional roles of inverted repeats.

Gene regulation. Negative supercoiling during transcription aids the formation of cruciforms which in turn have regulatory roles in transcription and transcription factor binding (Dayn et al. 1992), (Krasilnikov et al. 1999), (Branzei and Foiani 2010). Perfect palindromes are enriched upstream of translation start sites (Lu et al. 2007) and can be involved in alternative termination of transcription (Li et al. 1997). Finally, inverted repeats in introns affect alternative splicing of exons (Baraniak et al. 2003), (McAlinden et al. 2005).

Replication and recombination. Inverted repeats are non-uniformly distributed in relation to replication timing domains and their density increases from early to late replicating regions (Zou et al. 2017). Secondary structure formation at inverted repeats has been shown to result in replication stalling and interference with DNA polymerase progression (Voineagu et al. 2008). Additionally, long inverted repeats are more likely to stall the RNA polymerase (Lai et al. 2016). Inverted repeats have been also examined in viruses; and their role in initiation of replication has been described in detail, including description of inverted repeats in viral clusters (Pearson et al. 1996), (Chew et al. 2005), (Leung et al. 2005). Interestingly, an imperfect, long viral inverted repeat if mutated to a perfect inverted repeat results in extreme reduction of the replication rate to 5% of the original (Costello

et al. 1995). Application of monoclonal and polyclonal cruciform-specific antibodies has been shown to increase substantially the rate of replication (Zannis-Hadjopoulos et al. 1988). Similarly, a palindrome is located in the origin of replication in SV40 virus and exhibits helical instability contributing to the initiation of replication (Lin and Kowalski 1994). Additional roles have been indicated in intra- and inter- chromosomal recombination (Gordenin et al. 1993), (Tran et al. 1997), (Lobachev et al. 1998) as shown in yeast. Studies focusing on recombination generated mice with transgenes further indicated the role of palindromes in recombination in a mammalian model system (Akgün et al. 1997). However, further bioinformatic analyses would be required to measure the frequency of inverted repeats at replication initiation and recombination sites across organisms systematically.

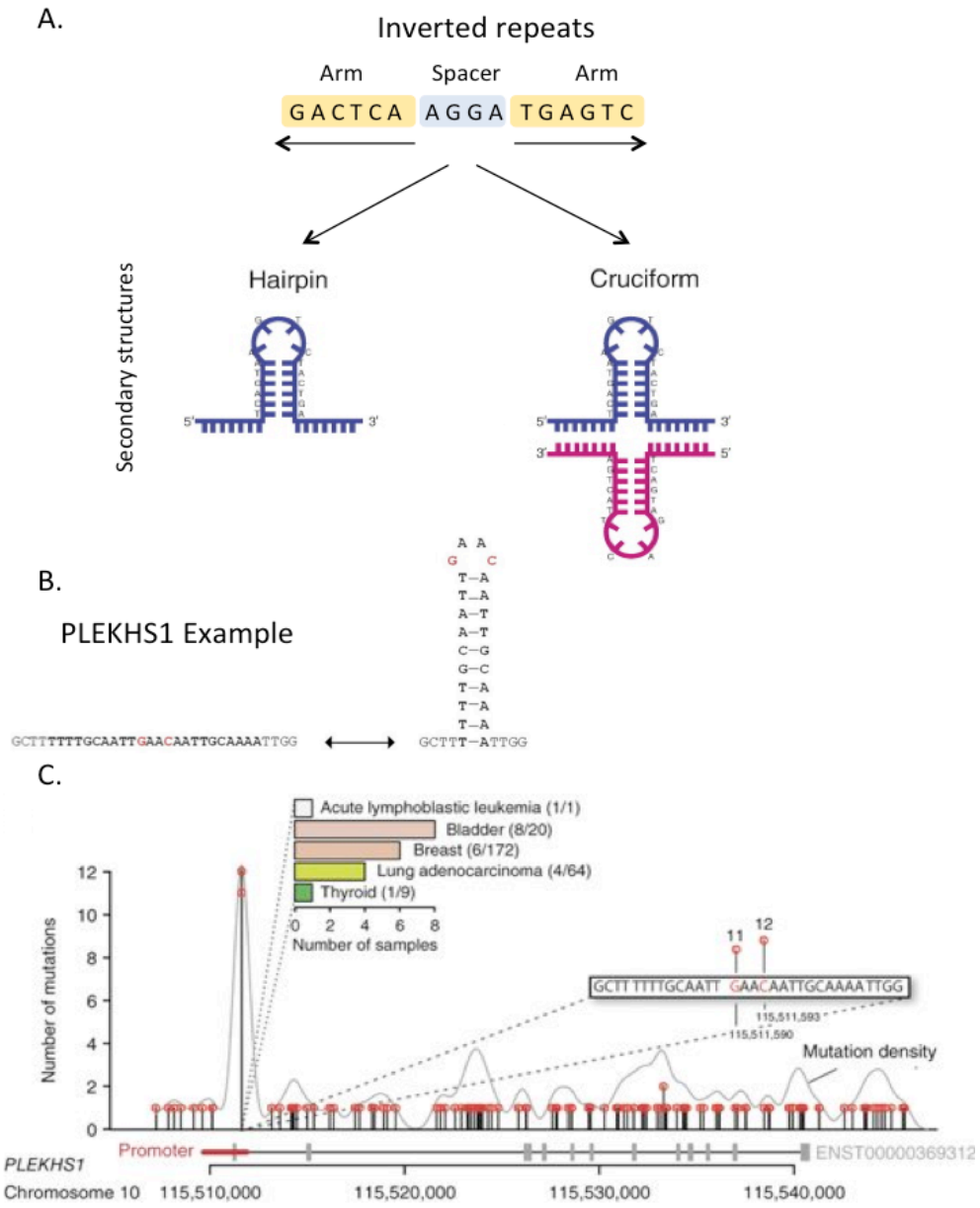


Figure 1.3: Secondary structure formation at inverted repeats and its effects on mutagenesis.

a. Inverted repeats are composed of two arms and an invervening spacer sequence. They can fold in hairpins and cruciforms. Adjusted from (Boyer et al. 2013). b. Inverted repeat at *PLEKHS1* gene promoter. Recurrently mutated sites are colored in red. Schematic from (Weinhold et al. 2014). c. Number of mutations across different cancer types overlapping the inverted repeat found in *PLEKHS1* gene promoter at two positions within its spacer sequence. The background rate of mutagenesis at the *PLEKHS1* genomic vicinity is also shown. Schematic from (Weinhold et al. 2014).

Mirror repeats and H-DNA.

A mirror repeat is composed of two adjacent copies of the same sequence, one of which is found in the reverse orientation. Similar to inverted repeats, they are composed of the two arms and the intervening spacer sequence which does not show the symmetry property (Figure 1.4a). They are also categorised into perfect and imperfect mirror repeats, the later of which has disruptions and mismatches in the arms. Up until now only a subset of mirror repeats have been shown to form secondary structures (Htun and Dahlberg 1988), (Wells et al. 1988). More specifically, AG / TC -rich mirror repeats have been shown to fold in intramolecular triple-stranded DNA, also known as H-DNA, in which a strand of DNA with mirror symmetry folds back to itself (Lyamichev et al. 1986), (Mirkin et al. 1987), (Frank-Kamenetskii and Mirkin 1995), (Figure 1.4b). Intermolecular triplexes can be generated either in normal cellular conditions or by using synthetic oligonucleotides, termed triplex forming oligonucleotides (TFOs) that can bind the DNA; these are also described below.

Intramolecular triple stranded DNA (H-DNA) forms at homopurine-homopyrimidine stretches that have mirror symmetry within them and they are thus a subset of mirror repeats. The third strand joins the double-stranded DNA and is held with Hoogsteen or reverse-Hoogsteen bonds while one strand remains single stranded. The result is a triple helical structure (Figure 1.4b). H-DNA can be classified depending on the nucleotide composition of the third strand, into pyrimidine-rich that bind with a parallel orientation with respect to the central strand or purine-rich in which case they bind in antiparallel orientation. The properties that influence the stability of H-DNA, including the arm length and nucleotide composition have been previously studied in detail in experimental models (Voloshin et al. 1988), (Frank-Kamenetskii and Mirkin 1995). For instance, parallel H-DNA is more stable than antiparallel H-DNA.

Evidence for H-DNA formation. Identification of the triplex structure was first produced by (Felsenfeld et al. 1957). However, evidence for triple-stranded DNA formation in biologically relevant models was first derived from analysis of supercoiled plasmids using

gel electrophoresis and Sl-nuclease hypersensitivity assays (Lyamichev et al. 1986), (Mirkin et al. 1987), (Frank-Kamenetskii and Mirkin 1995). Next, a major breakthrough was the development of antibodies against triple stranded DNA (Agazie et al. 1994), (Agazie et al. 1996). Fluorescent labelled oligonucleotides have been used to observe triple stranded DNA *in vivo* (Ohno et al. 2002). Moreover, structural evidence for its conformation dynamics has been generated with NMR spectroscopy (Koshlap et al. 1997). Case studies of H-DNA formation at particular genomic sites, with the most prominent H-DNA site found at the *c-MYC* promoter, have provided further evidence for their formation and their biological roles (Davis et al. 1989), (Wang et al. 2004), (Zhao et al. 2018). Furthermore, in certain promoters both G-quadruplexes and H-DNA structures can be formed by the same primary nucleotide sequence and it is unclear which mechanisms lead to preferences in one structure formation over the other. Also, environmental conditions influence the likelihood of H-DNA formation; with preference for low pH conditions (Lyamichev et al, 1985), (Mirkin et al. 1987). Biophysical studies have shown that triplex structure formation occurs at a substantially slower rate than for duplexes, with approximately 3 orders of magnitude difference (Rougée et al. 1992), (Sarai et al. 1993), (James et al. 2003), (Rusling et al. 2008), (Rusling et al. 2009).

Intermolecular triplexes. Intermolecular triplexes can occur between two DNA or RNA molecules or as a hybrid involving a DNA and an RNA molecule. In all of those cases Hoogsteen or reversed Hoogsteen bonds hold the structure together. If the third strand is an independent oligonucleotide molecule it is referred to as a triplex forming oligonucleotide (TFO). Evidence for intermolecular triplex formation commonly involves gel retardation assays, UV absorbance and CD spectroscopy experiments (Mortimer et al. 2014), (Zahran et al. 2015), (Kubota et al. 2015). Once they form, DNA-RNA triplexes are more stable than intramolecular DNA triplexes (Roberts and Crothers 1992).

Before Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technology was invented, triplex technology had received a lot of attention because oligonucleotides targeted at TFO binding sites could alter expression levels of the associated gene (Cooney et al. 1988), (Wu et al. 2008), (Rusling et al. 2008). They were also described

as potential therapeutic targets since they could generate site-specific mutagenesis (Moser and Dervan 1987), (Wang et al. 1996), (Vasquez et al. 2000), (Christensen et al. 2006). Nevertheless, with the advent of new technologies interest in TFO therapeutic applications has declined in recent years.

A number of different algorithms have been developed to identify intramolecular and intermolecular triplex DNA sites in the genome (Jenjaroenpun and Kuznetsov 2009), (Buske et al. 2012), (Buske et al. 2013), (Cer et al. 2013), (Hanzelmann et al. 2015). In addition, H-DNA sequences are highly enriched in eukaryotic genomes (Cox and Mirkin 1997) and Triplex Forming Sequence (TFS) sites have received considerable recent attention due to their potential to form DNA-RNA triplexes (Li et al. 2016).

Functional roles of mirror repeats, H-DNA and intermolecular triplexes.
H-DNA motifs are found with a frequency of approximately 1 every 50,000 nucleotides in the human genome (Schroth and Ho 1995). Nevertheless, their distribution in the human genome is variable. H-DNA sequences are enriched in introns (Bacolla et al. 2006) and promoters (Jain et al. 2008).

Gene regulation. H-DNA motifs are usually AG-rich. Since certain H-DNA motifs can fold into G-quadruplex structures, the thermodynamics of the folding between the two structures have also been investigated systematically. For instance, in the promoter of the oncogene *c-MYC*, the formed structure has been suggested to be both G-quadruplex and H-DNA but which conditions favour the formation of each of them remains unclear (Belotserkovskii et al. 2007), (Sun and Hurley 2009), (Zaytseva and Quinn 2018). However, formation of H-DNA in the *c-MYC* promoter has been shown to modulate its expression levels (Belotserkovskii et al. 2007).

H-DNA structures can be resolved by particular proteins. More specifically, the helicases WRN (Bacolla et al. 2011), DHX9 (Jain et al. 2013) and ChlR1 (Guo et al. 2015) have all been associated with unwinding of H-DNA structures *in vitro*. GAA repeats have been found to be linked to Friedreich's ataxia (Campuzano et al. 1996) and interestingly

evidence suggests that these repeats can form H-DNA structures (Frank-Kamenetskii and Mirkin 1995).

Triplex-forming oligonucleotides (TFOs) and Triplex target sequences (TTSs).
TFOs can module the expression of targeted genes (Rogers et al. 2005), (Hewett et al. 2006) and as a result have been suggested as potential cancer therapeutics. Triplex target sequences are highly enriched at regulatory regions and most profoundly at promoters, which is suggestive of regulatory functions of triplexes at those sites (Goñi et al. 2004), (Goñi et al. 2006).

DNA-RNA triplexes have roles in the recognition of chromatin by long non-coding RNAs (lncRNAs) (Li et al. 2016). For instance, *in vitro* pull-down assays and *in vivo* triplex-capture assays have demonstrated the formation of DNA-RNA triplexes between *Khps1* lncRNA and *SPHK1* promoter (Postepska-Igielska et al. 2015). Another study found DNA-RNA triplex forming sites to be enriched at promoters and introns in the human genome, with the stronger enrichment being at promoters (Jalali et al. 2017). Finally, a number of well characterised lncRNAs have been shown to form triplex structures including *HOTAIR*, *MEG3* and *FENDRR* among others (Grote and Herrmann 2013), (Mondal et al. 2015), (Kalwa et al. 2016). These results provide evidence for multiple functional roles of intermolecular triplexes in the cell.

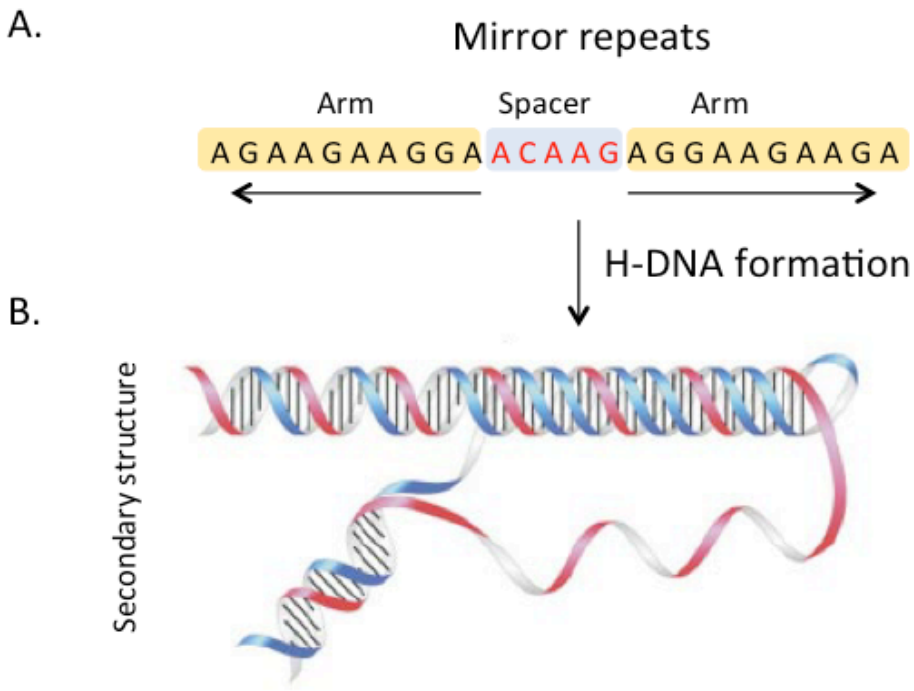


Figure 1.4: Formation of intramolecular triple-stranded DNA (H-DNA) at mirror repeat sequences.

a. Sequence with mirror symmetry and high AG content that has the potential to form H-DNA. b. H-DNA formation. Schematic of B. from (Wells 2008).

Direct repeats and short tandem repeats.

Direct repeats are composed of two identical sequences (arms) with a spacer sequence in between. In direct repeats, one repeat unit can misalign with the second repeat unit on the other strand, therefore generating a slipped structure that remains single stranded and exposed (Sinden et al. 2007), (Figure 1.5a). Short tandem repeats, also known as microsatellites, are defined as 1-9nt unit sequences, repeated at least 3 times, with a minimum size of ten nucleotides. There are over one million tandem repeats in the human genome and many of these are polymorphic. Similarly, to direct repeats, short tandem repeat units can misalign therefore forming slipped structures.

The first polymorphic microsatellite was reported by (Wyman and White 1980). PCR amplification has been traditionally used to measure the number of repeating units of tandem repeats present in a sample (Kovtun et al. 2007). Advances in sequencing technology have allowed the systematic interrogation of these sites. In particular, short tandem repeats that do not exceed certain size limitations can be investigated with current sequencing technologies (Hannan 2018). However systematic errors are encountered at large repeat tracts which are usually enriched in centromeric and telomeric regions. As a result, data from these regions tend to be discarded and they remain poorly mapped. Genic repeat sequences have been implicated in numerous disorders and are thought to be causative in the majority of them and may be key regulators in numerous biological processes. Since they are generally more mutable than the background rate in the genome they often serve as informative variable sites between individuals, resulting in individualised phenotypic differences between them.

Evidence for slipped structure formation. During gel electrophoresis direct repeats and short tandem repeats have slower electrophoretic mobility on the gel than other DNA of the same size (Pearson and Sinden 1996), (Panigrahi et al. 2005). In addition, the difference in electrophoretic mobility of tandem repeats is exaggerated with increase in their size (Tam et al. 2003). Electron microscopy has been applied to identify the formation of slipped structures (Axford et al. 2013); it was observed that the length of the tandem repeat is correlated with its propensity to form slipped DNA. Experimental evidence suggests slipped structure formation without the requirement of supercoiling and is remarkably stable under physiological conditions (Pearson and Sinden 1996).

In addition to slipped structures, other secondary structures can form at direct repeats and short tandem repeats. For instance, AC / GT short tandem repeats can also fold into Z-DNA, while a subset of GA / GAA short tandem repeats can fold into H-DNA (Bidichandani et al. 1998). Similarly, GGGN and GGGGN short tandem repeats can form G-quadruplexes. Hairpins can also form at trinucleotide repeat tracts such as $(CTG)_n$ in which case hydrogen bonding connects the two out of three base pairs in the repeating unit along the structure (Mitas 1997), (Darlow and Leach 1998). For R-loops and G-

quadruplexes, the stability of those structures might be altered by expansions and contractions of the repeat tracts (Santos-Pereira and Aguilera 2015). Additional evidence suggests that certain short tandem repeats can lead to the formation of Z-DNA at promoters, with regulatory roles (Rothenburg et al. 2001).

Finally, a multitude of methods and programs have been developed to identify short tandem repeats and direct repeats (Benson 1999), (Kolpakov et al. 2003), (Pokrzywa and Polanksi 2010), (Cer et al. 2013), (Lee et al. 2015), (Beier et al. 2017).

Functional roles of direct repeat and short tandem repeats.
Direct repeats and short tandem repeats tend to be highly polymorphic. As a result, it has been proposed that tandem repeats account for approximately 10-15% of the variance in gene expression (Gymrek et al. 2016). As sequencing technologies advance and the number of available genomes increases our understanding of the roles of those repeats in a number of different processes has continued to expand.

Gene regulation. Short tandem repeats are enriched in promoters (Sawaya et al. 2012), (Sawaya et al. 2013) and enhancers (Yáñez-Cuna et al. 2014) and have variable regulatory roles on gene expression (Kennedy et al. 1995), (Gebhardt et al. 1999), (Shimajiri et al. 1999), (Gymrek et al. 2016), (Quilez et al. 2016). In particular, 10-20% of eukaryotic genes and promoters contain an unstable repeat tract (Gemayel et al. 2010). Changes in repeat length at short tandem repeats have also been shown to be causal expression Quantitative Trait Loci (eQTLs) for expression variation (Borel et al. 2012). More specifically, polymorphic short tandem and direct repeats at promoters can have a strong effect on gene expression changes (eQTL analysis) and in methylation (mQTL analysis) in humans (Quilez et al. 2016), while presence of tandem repeats in the promoter is linked with higher variation in gene expression between individuals (Quilez et al. 2016).

Centromeres. Although tandem repeats are prevalent at centromeric sites, systematic investigation has remained extremely challenging with current sequencing technology.

As a result, advances in read length need to take place in order for those regions to be mapped accurately and examined for morphological and pathogenic characteristics. Therefore, their roles in centromeres remain poorly understood.

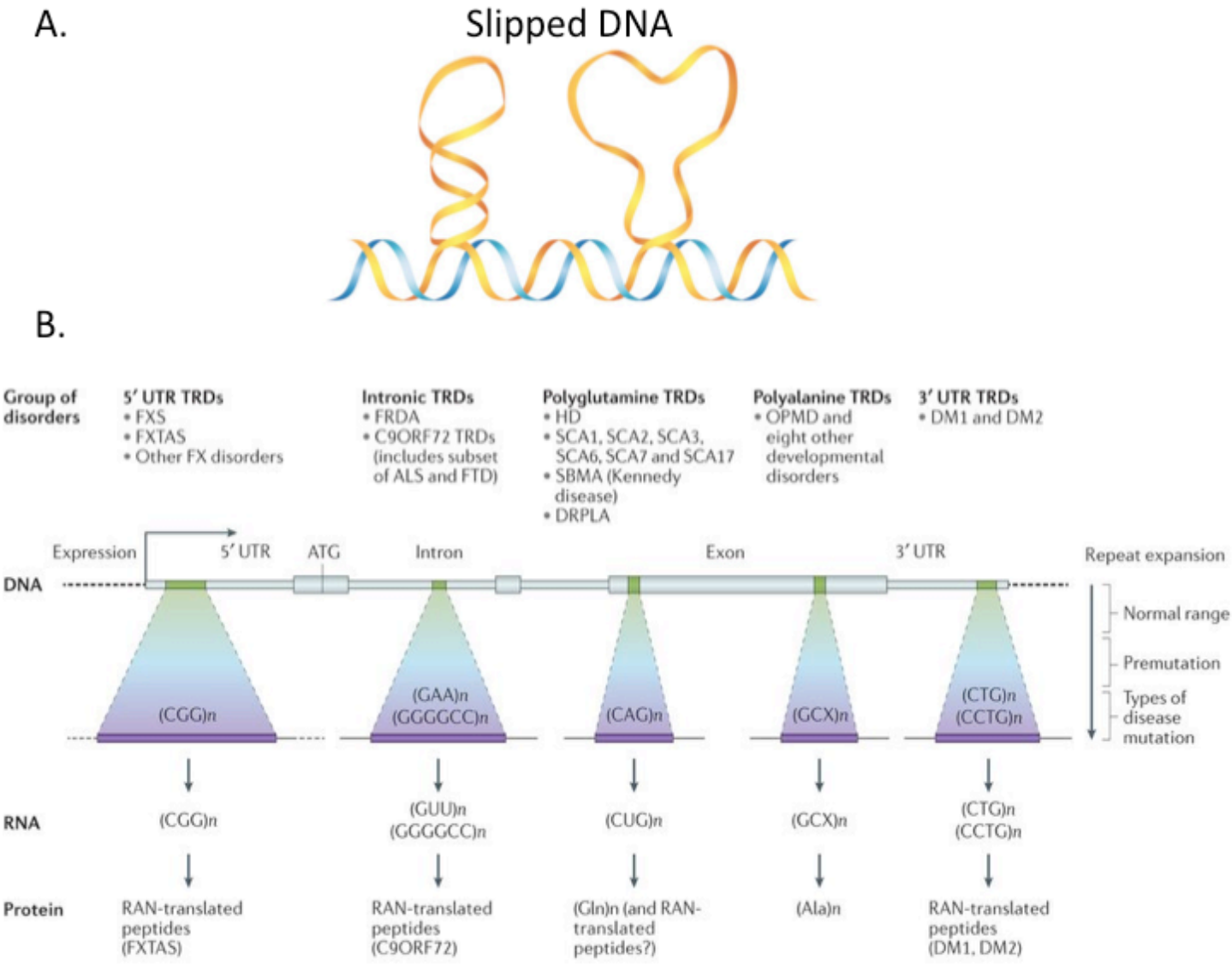


Figure 1.5: Slipped structures are directly implicated in multiple human disorders.

a. Slipped structure formation. Schematic from (Zhao et al. 2010). b. Multiple human disorders are the result of expansions and contractions of tandem repeat sequences throughout the gene length. Schematic from (Hannan et al. 2018).

Z-DNA.

Z-DNA is a left-handed double helical structure that is formed by alternating purine pyrimidine tracts (Wang et al. 1979), (Gessner et al. 1989), (Figure 1.6). In contrast to B-DNA, which has the *anti* conformation, in which atoms that hydrogen bond point away from the sugar, Z-DNA has a *syn* conformation. Z-DNA is less energetically favourable than B-DNA under physiological conditions; as a result, it requires negative supercoiling and can form during biological processes acting in the cell, such as transcription (Peck et al. 1982). Z-DNA is most easily formed at GC repeats, followed by GT repeats. It can also form at sites that do not conform to the alternating purine pyrimidine rule (Feigon et al. 1985). Similar to Z-DNA, formation of the same structure at the RNA level has been previously reported, and termed Z-RNA (Hall et al. 1984).

Experimental evidence for Z-DNA formation. Studies have provided both *in vitro* and *in vivo* evidence for formation of Z-DNA structures (Wang et al. 1979), (Singleton et al. 1982), (Peck and Wang 1983), (Haniford and Pulleyblank 1983a), (Haniford and Pulleyblank 1983b), (Nordheim and Rich 1983), (Jaworski et al. 1987), (Rahmouni et al. 1989). More specifically, antibodies have been raised against Z-DNA (Lafer et al. 1981), (Möller et al. 1982) and Z-DNA was subsequently detected *in vitro* (Nordheim et al. 1982). In the ciliate *Stylonychia mytilus* antibodies against Z-DNA stained macronuclei but not the micronuclei during sexual reproduction *in vivo*, suggesting that biological processes could mediate their formation by yet unknown mechanisms (Lipps et al. 1983). Case studies focusing on Z-DNA formation at the *c-MYC* promoter provided evidence that transcription mediated its formation and by switching off transcription, Z-DNA formation was quickly halted (Wittig et al. 1992), (Wölfl et al. 1995), (Wölfl et al. 1996).

Nucleotide composition and disruptions in the sequence of Z-DNA structures have also been analysed extensively (Ellison et al. 1985). Additionally, the region where the transition between B-DNA and Z-DNA occurs also known as the B-Z junction has been investigated in depth using biophysical models (Soumpasis et al. 1987), (Doktycz et al. 1990) and its crystal structure has been resolved (Ha et al. 2005). During Z-DNA

formation, in the B-Z junction two bases are extruded which are exposed and prone to enzymatic or chemical alterations (Ha et al. 2005). Additionally, Z-DNA can be stabilised by chemical compounds, including spermine and spermidine (Thomas et al. 1991), while its formation is also promoted by methylation (Zacharias et al. 1990).

Z-DNA interacts with a number of proteins. Methods have been developed to isolate proteins that bind preferentially to Z-DNA from those that are bound to B-DNA (Herbert and Rich 1993) and found ADAR1 protein as one of the proteins that binds strongly and specifically to Z-DNA (Herbert et al. 1995). Also, multiple proteins that have a Z-DNA binding domain have been identified (Herbert et al. 1998), (Kim et al. 1999), (Schwartz et al. 1999). Using the specificity of these proteins ChiP-seq experiments have been performed to identify sites of Z-DNA formation in the genome (Shin et al. 2016).

A number of different computational programs have been developed to identify sequences that can form Z-DNA (Schroth et al.1992), (Champ et al. 2004), (Li et al. 2009), (Cer et al. 2013).

Functional roles of Z-DNA.

Z-DNA sequences occur at an average rate of 1 every 3,000 nucleotides in the human genome (Khuu et al. 2007). Nevertheless, the distribution of Z-DNA sequences in the human genome is non-uniform with strong enrichment close to transcription start sites (Schroth et al. 1992) in accordance with the negative supercoiling found there, which mediates their formation.

Gene-regulation. Z-DNA forms readily at promoters of transcribed genes due to negative supercoiling (Wölfl et al. 1995) and inhibition of transcription results in a reduction of Z-DNA formation as measured by antibodies that are specific to Z-DNA (Wittig et al. 1991), (Wölfl et al. 1996). Z-DNA sequences are enriched around the transcription start sites of multiple human genes (Schroth et al. 1992), (Li et al. 2009). Multiple studies of the *c-MYC* promoter have described in great detail the formation of multiple structures, including Z-DNA, which was found to form in three separate parts of the promoter and regulated its

transcriptional levels (Wittig et al. 1992). In support to that, formation of Z-DNA at the *CSF1* gene, aids the recruitment of the RNA polymerase after stabilising the open chromatin (Liu et al. 2001).

Similarly, to formation of G-quadruplexes, Z-DNA does not allow the formation of nucleosomes and could therefore stabilise transcriptional processes that require open chromatin (Garner et al. 1987), (Wong et al. 2007). For instance, transcription factor binding could follow the formation of Z-DNA and nucleosome displacement (Rich and Zhang 2003). On the other hand, Z-DNA forming in *ADAM-12* first exon has been investigated in detail and it was found that it acts as a transcriptional repressor element (Ray et al. 2011), (Ray et al. 2013). However, further investigation is required to fully characterise the functional roles of Z-DNA upstream and downstream of the transcription start site.

Figure 1.6: Formation of left-handed Z-DNA.

Schematic from (Bagshaw 2017).

1.2.2. Relationship between non-B DNA and mutability.

Mutability across the human genome is unevenly distributed. Non-B DNA secondary structures have been previously associated with genomic integrity (Bacolla and Wells 2004), (Wang and Vasquez 2006), (Raghavan et al. 2006), (Zhao et al. 2010). Indeed, early experimental investigations of older mice showed an excess of single stranded DNA formation in multiple tissues (Price et al. 1971), (Chetsanga et al. 1975), (Nakanishi et al. 1979); this is of interest since single stranded DNA formation has been used as a proxy

for secondary structure formation (Kouzine et al. 2017). Moreover, there is accumulating evidence for causal implication of non-B DNA motifs in cancer mutagenesis. Below, the relationship between each non-B DNA motif and different types of mutations is described.

G-quadruplex.

G-quadruplex formation is elevated in cancer cells in comparison to normal cells, as shown using G-quadruplex specific antibodies coupled with immunohistochemistry (Biffi et al. 2014). In combination with their enrichment at telomeres, regulatory regions and oncogenes, G-quadruplexes have received considerable attention as an actionable target for potential anti-cancer treatments (Hänsel-Hertsch et al. 2017). Nevertheless, G-quadruplexes are not only associated with cancer, there is evidence that suggests that they are also implicated in mutagenesis.

G-quadruplexes are inherently more mutable than their surrounding sequences and population studies find genetic variants enriched at G-quadruplex sequences (Du et al. 2014). Other studies have provided evidence for enrichment of G-quadruplexes at sites of copy number variation (CNV) across multiple human cancers (De and Michor 2011). Interestingly, eQTLs are also more likely to fall within G-quadruplexes than in surrounding sequences, but the enrichment is smaller than the observed enrichment across all genetic variants (Du et al. 2014). Nevertheless, this suggests that they could harbour a pool of variants that contribute to expression changes by unexplored or poorly understood processes.

In addition, given the non-random distribution of G-quadruplexes in the human genome and their enrichment at regulatory sites, it remains plausible that mutations overlapping them could have functional consequences that have not been characterised. In support to that, G-quadruplexes are found at the site of breakpoint formation and could contribute to chromosomal fragility (Nambiar et al. 2010). Additionally, comprehensive case studies of the oncogenes *C-KIT*, *KRAS*, *BCL2* and *c-MYC* indicate that G-quadruplexes in their

promoters regulate expression levels (Siddiqui-Jain et al. 2002), (Rankin et al. 2005), (Cogoi et al. 2006), (Dai et al. 2006), (Phan et al. 2007), (Morgan et al. 2016). Indeed, a mutation at a single nucleotide (G->A) of a G-quadruplex that functions as a transcriptional repressor at *c-MYC* promoter results in 3-fold increase in its expression levels, while the promoter expression levels are modulated by a G-quadruplex binding and stabilising ligand (Siddiqui-Jain et al. 2002).

Nevertheless, until now there haven't been systematic studies that examine how mutations overlapping G-quadruplexes at regulatory elements across cancer patients could affect the expression levels of cancer-associated genes and serve as driver mutations in cancer progression. In addition, their mutational profiles have not been analysed across cancer patient genomes and tumour types.

Inverted repeats.

Inverted repeats are intrinsically mutagenic. They have been associated with mutagenesis in a number of studies in diverse organisms, both in prokaryotic and in eukaryotic systems for different mutation types (Gordenin et al. 1993), (Nag and Kurst 1997), (Lobachev et al. 1998), (Echlin-Bell et al. 2003), (Lobachev et al. 2007), (Eykelenboom et al. 2008), (Du et al. 2014), (Lu et al. 2015), (Kamat et al. 2016), (Nik-Zainal et al. 2016), (Zou et al. 2017). In human cancers, substitutions, indels and rearrangements have been related to inverted repeats through hairpin or cruciform structure formation. In a recent systematic experimental study, short inverted repeats were found to be enriched for somatic mutations and at hotspots of genetic instability (Lu et al. 2015). More specifically, they were enriched in substitutions, indels and breakpoints of rearrangements (Lu et al. 2015).

Early studies demonstrated that inverted repeats induce double strand breaks resulting in indels in prokaryotic cells (Collins 1981), (Collins et al. 1982). *E. coli* cells when introduced with a long palindromic inverted repeat can replicate when recombination is

present but not in recombination deficient cells (Leach et al. 1997). This is the result of double strand break formations at inverted repeats that cannot be effectively repaired in the mutant cells. In addition, hairpins as well as other non-canonical DNA structures can cause replication fork stalling both in prokaryotic and eukaryotic cells (Samadashwily et al. 1997), (Voineagu et al. 2008). Similarly, recent studies have demonstrated that in eukaryotic cells inverted repeats can induce double strand break formation and can stall DNA replication fork progression *in vivo* (Lu et al. 2015). In particular, the influence of biophysical properties of inverted repeats have also been investigated; deletions and recombinations at inverted repeats are correlated to their length and anti-correlated to the size of the spacer sequence of the inverted repeat (Lobachev et al. 1998).

Recurrent translocations have been found in AT-rich inverted repeats (Kurahashi et al. 2006) and in particular at cruciform-forming inverted repeats (Gotter et al. 2004), (Inagaki et al. 2009). Additionally, there is evidence for the role of imperfect repeats, which can contain mismatches in the arms, in indel mutagenesis (de Boer and Ripley 1984), (Rosche et al. 1997). Moreover, substitutions were found overlapping inverted repeats and creating perfect inverted repeats from imperfect by mutations in their arms and are also implicated in the generation of missense and nonsense mutations (Kamat et al. 2016).

Recent analysis of cancer genomes has expanded our understanding for the role of inverted repeats in mutagenesis. Recurrent mutations are enriched in a group of inverted repeats, the most prominent being an inverted repeat at *PLEKHS1* promoter, mutated at two sites recurrently for different cancer types across multiple patients (Weinhold et al. 2014), (Nik-Zainal et al. 2016), (Zou et al. 2017), (Figure 1.3b-c). The recurrent mutations are more enriched at the spacer sequence than in the arms, which is more exposed and thus more likely to mutate. Additionally, the subset of inverted repeats with the spacer sequence motifs “GAAC” and “GTTC” are much more mutable, raising the possibility that they are also more likely to cause hairpin formation and induce mutagenesis (Zou et al. 2017). In support to that argument, biophysical studies have indicated that inverted repeats with spacer sequence of 4-6nt are more likely to form hairpin structures

(Rentzeperis et al. 1993). Further evidence comes from genome-wide analysis of inverted repeats with the “GAAC” and “GTTC” spacer sequence, which were found recurrently mutated in cancer patients in multiple loci (Nik-Zainal et al. 2016). Therefore, recurrent hypermutation of inverted repeats with a particular spacer motif occurs not only at the *PLEKHS1* promoter but in a number of other sites. Nevertheless, evidence for changes in expression levels of *PLEKHS1* due to mutation of its inverted repeat is still lacking (Weinhold et al. 2014), (Nik-Zainal et al. 2016), (Zou et al. 2017), which could be suggestive of an increased likelihood of mutagenesis but no selective advantage gains associated with it. As a result, inverted repeats could be potential endogenous agents of DNA damage in the human genome.

Mirror repeats.

Substitutions and indels have been found to be enriched at mirror repeats, even those that do not fall in the H-DNA category (Cooper and Krawczak 1991), (Sinden and Wells 1992), (Kamat et al. 2016). Although the reason for their enrichment at non-HDNA motifs remains unclear, one possibility is that the symmetry property allows the formation of alternative conformations between the two strands in a yet unknown mechanism. One recently proposed mechanism is through the formation of a Moebius loop (Kamat et al. 2016) which has not yet been shown *in vitro* or i*n vivo*.

H-DNA.

Sequences that predispose to H-DNA formation have been associated with genomic instability in prokaryotic and eukaryotic organisms including human cells (Wang and Vasquez 2004), (Zhao et al. 2018). H-DNA is inherently mutagenic and recombinogenic (Rooney and Moore 1995), (Faruqi et al. 2000), (Wang and Vasquez 2004). Moreover,

ERCC1-XPF, XPG and FEN1 can directly cleave H-DNA structures and are associated with H-DNA associated mutability *in vivo* (Zhao et al. 2018).

Early evidence for its mutagenic role in human cells came from experimental studies focusing on H-DNA formation in the promoters of *c-MYC* and *BCL2*. A number of different studies examined H-DNA structure formation in the *c-MYC* promoter and found it to be enriched for rearrangements in multiple different cancers (Carè et al. 1986), (Mikrin et al. 1987), (Haluska et al. 1988), (Joos et al. 1992), (Saglio et al. 1993), (Wilda et al. 2004). Additional work showed that H-DNA formation at the *c-MYC* promoter causes double strand breaks, deletions and rearrangements (Wang and Vasquez 2004), (Wang et al. 2008). Furthermore, transgenic mice with inserted H-DNA structures were reported to have more chromosomal deletions and translocations at those regions (Wang et al. 2008), directly implicating these structures in mutagenesis. H-DNA formation in *c-MYC* has been shown to stall RNA polymerase progression both *in vitro* and *in vivo* providing mechanistic insight into its inherent mutability (Krasilnikova et al. 1998), (Belotserkovskii et al. 2007), (Krasilnikova et al. 2007). Similarly, the H-DNA structure found in *BCL2* promoter was found enriched for rearrangements (Raghavan et al. 2005), (Zhao et al. 2018). In particular, H-DNA formation at *BCL2* promoter overlapped the major breakpoint region of follicular lymphoma and its formation was indicated with antibodies raised against it (Raghavan et al. 2005). In support of that, a three-nucleotide mutation disrupting the same H-DNA sequence resulted in reduction in rearrangements at the site.

Systematic investigation has provided further evidence that H-DNA sequences are enriched for rearrangements (Bacolla et al. 2006). Additionally, indels and substitutions were also shown to be enriched at H-DNA motifs (Kamat et al. 2016). Finally, H-DNA sequences accounted for about 5% of microinsertions and microdeletions in the same study (Kamat et al. 2016).

Intermolecular triplexes have also been implicated in mutagenesis. In fact, site specific mutagenesis by designed oligos can occur at triplex target sites (Vasquez and Wilson 1998), (Vasquez et al. 2000). This finding had received attention a decade ago to be used

for gene editing and gene therapies due to its potential for site-specific mutagenesis. Nevertheless, the advance of TALEN and later CRISPR has provided new technologies with higher efficiency and ease of use and the interest has since declined.

Direct repeats and short tandem repeats.

Short tandem repeats and direct repeats are some of the most mutable sequences in the genome. Short tandem repeats have been implicated in a number of disorders including a number of Mendelian monogenic and developmental disorders and multiple trinucleotide repeat disorders, many of which have been characterised in depth. More specifically, short tandem repeats have been found responsible for more than 30 Mendelian disorders (Mirkin 2007), (Figure 1.5b). These include Amyotrophic lateral sclerosis (ALS), Fragile-X Syndrome and Huntington's disease among others.

For example, in ALS, the repeat GGGGCC in the gene *C9ORF72* expands in size in ALS patients (DeJesus-Hernandez et al. 2011), (Renton et al. 2011). In Fragile X Syndrome an expansion of a short tandem repeat at the 5'UTR of *FMR1* is the cause of the disorder. *FMR1* alleles with 55-200 CGG•CCG-repeats have been shown to result in neurodegeneration (Hagerman and Hagerman 2004). Similarly, in Huntington disease a CAG repeat in *Huntingtin* gene is found to have increased in the number of its copies present which causes neurodegeneration. Other disorders include Myotonic Dystrophy, Friedreich's Ataxia, Spinocerebellar ataxias and Frontotemporal dementia among many. Therefore, a number of well characterised disorders are mediated by the expansion or contraction of tandem repeats in genic regions (Figure 1.5b).

Importantly, it is thought that tandem repeats impact is not limited to Mendelian, monogenic disorders. They also have roles in polygenic disorders that are currently not fully understood (Gymrek et al, 2016), (Quilez et al. 2016). A bold suggestion has been that short tandem repeats actually account for the missing heritability which currently cannot be identified by standard methods, measuring heritability for polygenic disorders

(Hannan 2010), (Press et al. 2014), (Gymrek et al. 2016). Indeed, one in twenty proteins contains at least one polymorphism in an unstable tandem repeat (O'Dushlaine and Shields 2008), while similar polymorphisms in non-coding regions remain less explored, especially in regions that remain difficult to map accurately.

In cancer, short tandem repeats have been found to be hypermutable in a subset of patients that have mismatch repair deficiency. Mismatch repair deficient cancers show microsatellite instability (MSI) (Vilar and Gruber 2010) and the Bethesda Panel of short tandem repeat markers or immunohistochemistry staining biopsy tests have been used for classification in high level microsatellite instability (MSI-H), low level microsatellite instability (MSI-L) and microsatellite stable (MSS) tumours (Murphy et al. 2006), (Kim et al. 2013). A clinical surprise has been the observation that MSI patients have a higher survival likelihood than MSS patients. This is likely because of the excess of mutations in mismatch repair deficient patient tumours which can exceed tens of thousands to even hundreds of thousands of somatic mutations in a single cancer genome or because random mutations can trigger neoantigen generation that is 'non-self' and can be targeted by the immune system; therefore, being more responsive to immunotherapy (Le et al. 2015). The frequency of MSI-H samples varies by tumour type; colorectal carcinomas are one of the most frequent mismatch repair deficient tumour types and have about 15% of samples being MSI-H (Gatalica et al. 2016). The observation that MSI patients have a higher survival likelihood has been critical for targeted treatments (Hewish et al. 2010). Interestingly, mutational signatures can be used to accurately predict MSI status (Davies et al. 2017).

Minisatellites have been used in genotoxicity studies to measure the levels of ionising radiation (Dubrova et al. 1993), (Dubrova et al. 1998). Additionally, transcription factor binding sites can be amplified or deleted at polymorphic satellites. For instance, the EWS-FLI1 transcription factor which drives Ewing sarcoma binds to the GGAA repeat (Riggi et al. 2014), suggesting that mutations of those sequences could have functional effects. Interestingly, Ewing's sarcoma (EWS) transcription factor also binds at G-quadruplex sites (Takahama et al. 2011). Finally, satellite RNA is transcribed from DNA satellite

regions in the genome. Recent evidence suggests that BRCA1-deficient breast cancer tumours have an excess of satellite RNAs (Zhu et al. 2011), (Zhu et al. 2018). In addition, overexpression of satellite RNAs is tumorigenic in mice (Zhu et al. 2018) and mechanistically they can often associate in R-loops that destabilise the replication fork machinery.

Z-DNA.

Z-DNA is intrinsically mutagenic both *in vitro* and *in vivo* (Wang et al. 2008). In mammalian cells, Z-DNA sequences induced double strand breaks which resulted in large deletions in those cells (Wang et al. 2004). In contrast to that, in *E. coli* the same sequences resulted in smaller deletions (Collins 1981), (Freund et al. 1989), (Wang and Vasquez 2006), implicating differences in the repair processes in these two model systems. Furthermore, large Z-DNA sequence insertions in bacteria using plasmids were unstable, indicating their high mutagenic potential (Wang and Vasquez 2006). Furthermore, several reports have shown that Z-DNA sequences are enriched in chromosomal breakage hotspots across multiple different genes (Adachi and Tsujimoto 1990), (Rimokh et al. 1991). The large deletions and rearrangements associated with Z-DNA structure formation are not only found directly overlapping the Z-DNA motifs but also in the nearby vicinity from the Z-DNA site (Wang et al. 2006).

It has also been shown that transcription levels are correlated with the mutability of Z-DNA sequences in eukaryotic cells (Wang et al. 2006), and it has been suggested that Z-DNA structures could interfere with RNA polymerase progression. In further support to that claim, the level of mutagenesis due to Z-DNA structures in bacteria was linked to the level of expression of the corresponding genes (Jaworski et al. 1989), while the sites of chromosome breakage overlapping Z-DNA sequences tend to occur at promoters (Adachi and Tsujimoto 1990), (Rimokh et al. 1991), which could suggest negative supercoiling induces more frequent Z-DNA formation, which in turn could generate genomic instability. Similarly to formation of secondary structures at inverted repeats,

$(GC)_{14}$ tandem repeats, which could fold in Z-DNA formation were found to inhibit RNA polymerase progression; in contrast to that $(GC)_{14}$ control sequences with same length and GC content that did not display the alternating purine-pyrimidine property did not inhibit the RNA polymerase (Ditlevson et al. 2008). Additionally, DNA damage is harder to repair when DNA adopts the Z-DNA conformation (Lagravère et al. 1984), (Boiteux et al. 1985). Perhaps this is the result of DNA repair enzymes not being able to access the site of DNA damage during Z-DNA formation or other physical constraints. Finally, although the interactions between Z-DNA structures and several proteins have been investigated (Rich and Zhang 2003), the precise mechanisms that recognise and cleave Z-DNA structures are not yet known.

Aims of the thesis

- Characterise the distribution of non-B DNA motifs in the human genome.
- Investigate the relationship between non-B DNA motifs and somatic mutability (substitutions, insertions, deletions and rearrangements) across cancer genomes.
- Explore the relationship between non-B DNA motifs and recurrent mutagenesis across cancer patients.
- Construct genome-wide models of mutagenesis in the human genome using non-B DNA motifs, epigenetic markers and replication timing as predictive features.
- Produce a comprehensive analysis of genomic features influencing insertion and deletion mutagenesis in the human genome.
- Devise a method to estimate transcriptional and replicative strand asymmetries at indel sites across multiple cancer types.
- Develop mechanistic insight into the mutational processes that produce strand asymmetries at indel sites.

CHAPTER TWO

2. Genome-wide characterisation of non-B DNA motifs and somatic mutations.

In this chapter, each of the non-B DNA motif categories is introduced and characterised in relationship to its genomic properties such as length distributions and frequencies in the human genome. Next, the relative enrichment patterns of each non-B DNA motif are investigated across functional genomic sites and epigenomic features. Mutational enrichment patterns of non-B DNA motifs are estimated across ten cancer types, in support to their roles in cancer mutagenesis. More specifically, it is shown that non-B DNA motifs are more mutable than the surrounding regions contributing to locally elevated mutational rates in both substitution and indel mutagenesis.

2.1. Distribution in the genome and characteristics of non-B DNA motifs.

2.1.1. Non-B DNA motifs: their algorithmic identification in the genome.

In total, seven non-B DNA motif categories were used across the analysis, these included mirror repeats, H-DNA, short tandem repeats, Z-DNA, inverted repeats, direct repeats and G-quadruplexes (Figure 2.1) each of which was described in detail in chapter 1. The annotations for the genome-wide locations of the non-B DNA motifs in the human genome were derived from (Cer et al. 2013). A description of the non-B DNA motifs from an algorithmic perspective is described below and the conventions are followed throughout the thesis:

● A mirror repeat is composed of two adjacent copies of the same sequence, one of which is found in the reverse orientation. The minimum length of each of its arms is 10bp and the overall length of the motif is equal to or greater than 20bp. A subset of mirror repeats are termed Hinged DNA (H-DNA), because they are predisposed to forming an intramolecular triple helical structure connected through Hoogsteen bonds. H-DNA sequences have a high (>90%) AG content, arm lengths of >=10bp and spacer size of less than 8bp as described previously (Cer et al. 2013).

● Z-DNA is alternating purine-pyrimidine tracts of at least 12bp that can form a left-handed double helical structure. However, Z-DNA sequences cannot contain within them AT dinucleotides.

● Direct repeats are composed of two copies of the same sequence with a potential spacer sequence in between. They have a minimum length of 20bp and minimum arm length of 10bp. Direct repeats can misalign, therefore forming slipped structures.

● Short tandem repeats denote repeat sequences of unit size of 1-9bp. The minimum total length of the motifs is 9bp and the repeating unit is repeated three or more times. Short tandem repeats are prone to misalignment, replication slippage and formation of looped or slipped structures.

● Inverted repeats are composed of two adjacent copies of the same sequence, one of which is found in the reverse complement orientation. The minimum arm length is 6bp, and the spacer can range between 0 to 100bp. These sequences are capable of forming hairpin and cruciform structures.

● G-quadruplexes are defined with the consensus motif: $G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}N_{1-7}G_{\geq 3}$. G-runs denote the part of three or more consecutive guanines which can form Hoogsteen bonds, whereas the variable region (N) which forms the loops is the spacer sequence and remains single stranded.

For each non-B DNA motif, sequences that overlapped centromeric sites and assembly gaps and which result in common artefacts were filtered out. Additionally, low complexity regions, such as those found at telomeres, at which mutation calling with short reads using the Wellcome Sanger Institute pipeline is unreliable were also excluded. Regions of the genome with exceptionally high sequencing depth from the 1,000 Genomes Project (top 0.01% Hi Seq Depth) were also excluded (Pickrell et al. 2011). In particular, for short tandem repeats this resulted in the exclusion of >30% of their total occurrences, which cannot be examined with current sequencing technologies and the short reads implemented in whole genome sequencing of cancer patients.

In the next section the genome-wide maps of human non-B DNA motifs were used to investigate: i) the characteristics of each non-B DNA motif, ii) the amount of overlap between non-B DNA motifs, iii) potential relationships between each non-B DNA motif and other genomic and epigenomic features, iv) the distribution of non-B DNA motifs around functional elements at a nucleotide resolution.

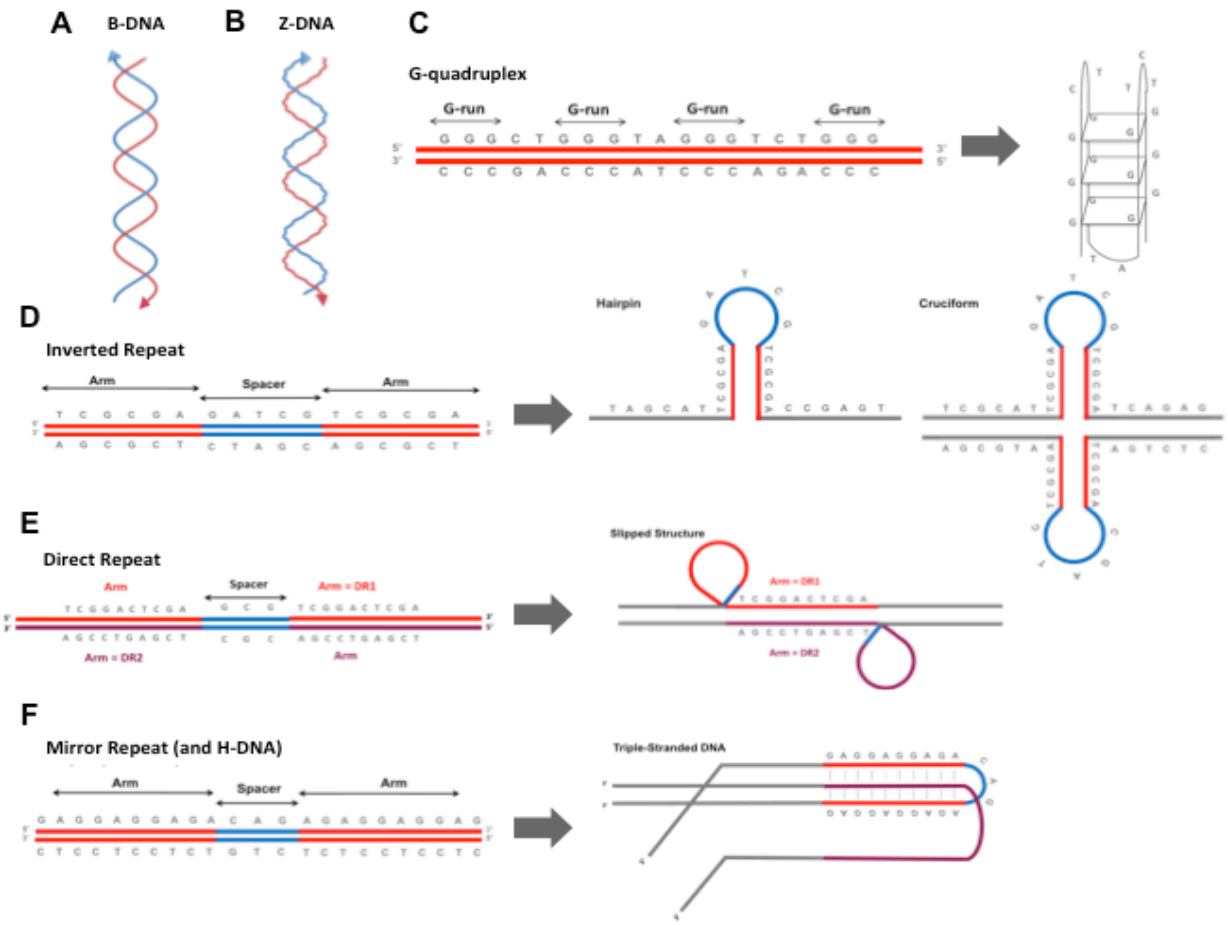


Figure 2.1: Non-canonical secondary structures arising from non-B DNA motifs in the human genome.

a. Normal configuration of human DNA. b. Left-handed helical structure caused by Z-DNA (c-f). Schematic representations of the primary sequence of various non-B motifs and their corresponding predicted secondary structures.

2.1.2. Characteristics of non-B DNA motifs.

The distribution and genome coverage for each non-B DNA motif was investigated. Each non-B DNA motif was found to account for approximately 0.07% to 4% of the mappable human genome (Figure 2.2a). The number of occurrences of individual non-B DNA motifs varied by almost two orders of magnitude and ranged between 69,154 to 6,006,266. The most frequent type of non-B DNA motif was inverted repeats and the rarest were H-DNA

motifs. Additionally, the length distribution of each non-B DNA motif category was investigated and it was found that the vast majority of non-B DNA motif occurrences were less than 50 nucleotides long (Figure 2.2b-c).

It was investigated whether different types of non-B DNA motifs aggregate in the human genome. Indeed, certain sequence constraints do not allow the overlap of particular non-B DNA motif pairs such as G-quadruplexes (which contain G-runs) and Z-DNA (which involves alternating purine-pyrimidine stretches) and would be expected to have very limited overlap. However, it has been noted that in certain cases such as H-DNA and G-quadruplex, both structures could be formed by the same primary sequence. For instance, in *c-MYC* promoter at the same site, H-DNA and G-quadruplex formation have both been reported previously (del Mundo et al. 2017). The Jaccard index is the intersection over the union of two features and here was implemented to measure the similarity of two sets of non-B DNA motifs and in particular to measure how often non-B DNA motifs overlap in the human genome over their union of occurrences. A Jaccard index of 1 would indicate that two non-B DNA motifs occur always together, whereas a Jaccard index of 0 would imply that the two non-B DNA motifs are always found separately in the human genome. Limited overlap was observed between non-B DNA motif categories (Figure 2.2d).

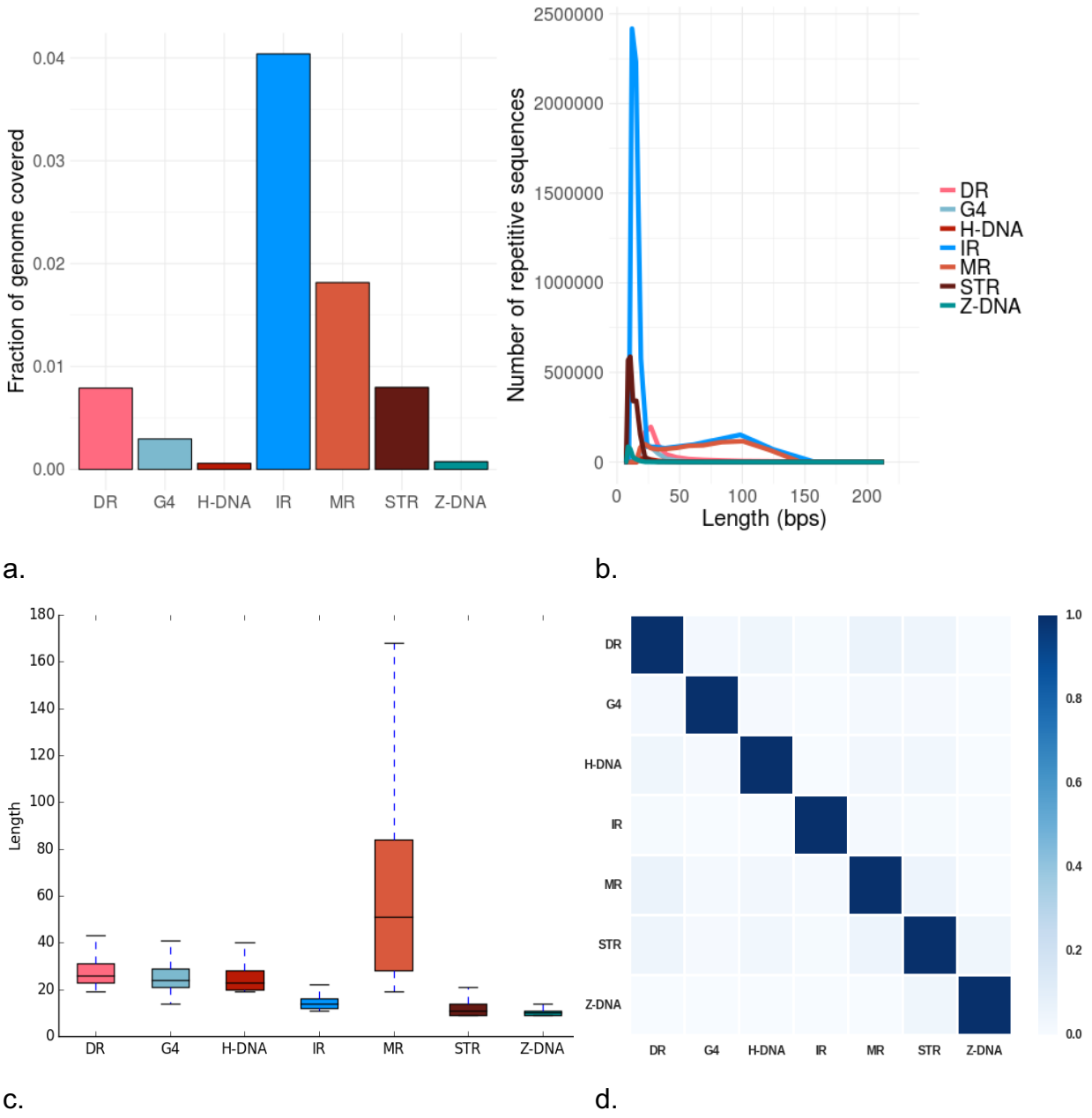


Figure 2.2: Genome properties of non-B DNA motifs.

a. Fraction of the human genome covered by different non-B DNA motifs. b. Distribution of number of non-B DNA motifs and their lengths. c. Boxplot of lengths for non-B DNA motifs. d. Jaccard index heatmap reporting the amount of overlap between different non-B DNA motifs. 1 represents complete overlap, whereas 0 represents no overlap.

2.1.3. Non-B DNA motifs and genomic partitions.

Next, a systematic investigation was performed regarding the localisation of each non-B DNA motif in the human reference genome, exploring their relationship with functional sites and histone modifications. In particular, two independent methods were used to investigate differential enrichment of each non-B DNA motif across chromatin states and functional elements of the human genome. Using both methods independently, it was shown that the distribution of non-B DNA motifs in the genome follows non-uniform patterns with certain non-B DNA motifs being particularly enriched at promoter regions.

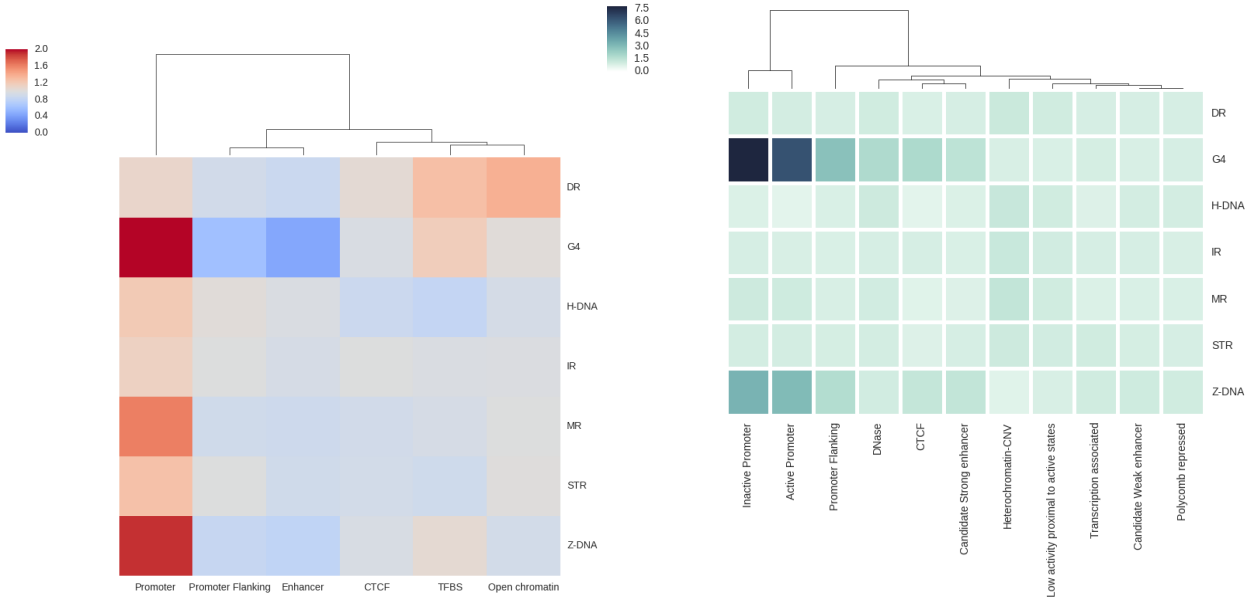
The first method that was employed was based on the Ensembl regulatory features, and utilises cell-type-independent chromatin annotations of regulatory regions (Zerbino et al. 2015). This method uses a genome segmentation algorithm based on known genomic features, experimentally-derived histone modifications and transcription factor binding sites to derive the regulatory features and is manually curated. Using this method an enrichment of all non-B DNA motif categories was observed at promoters relative to other functional elements, which varied by non-B DNA motif and was most prominent for Z-DNA and G-quadruplexes (Figure 2.3a). This enrichment is suggestive of diverse roles of multiple non-B DNA motifs in promoter regions, some of which have been previously explored by experimental studies (Rhodes and Lipps 2015), (Armas et al. 2017). For instance, G-quadruplex formation at promoter sequences has dramatic effects in expression changes (Lam et al. 2013), while negative supercoiling at promoters during transcription may induce the formation of DNA secondary structures (Ma and Wang 2016). A relative depletion of non-B DNA motifs was observed at enhancer regions, the levels of which varied by non-B DNA motif category (Figure 2.3a).

Similarly, a second method was used to further investigate these observations. The second method was based on a Bayesian network that generated chromatin state annotations (Segway) for six cell types from a group of genome-wide assays, including DNA-seq and ChIP-seq (Hoffman et al. 2012), (Hoffman et al. 2013). The Segway analysis validated the clear enrichment of Z-DNA and G-quadruplex sequences at GC-

rich promoter regions (Figure 2.3b). Also, it was found that direct repeats, H-DNA and mirror repeats are modestly enriched in low complexity, repetitive sequences (e.g. heterochromatin); and inverted repeats and short tandem repeats are uniformly distributed between gene-rich through to gene-poor regions (Figure 2.3b). Note, this analysis excluded low mappability repeat regions of the genome, which cannot be investigated accurately with current sequencing technologies and for which mutation calling would not be reliable.

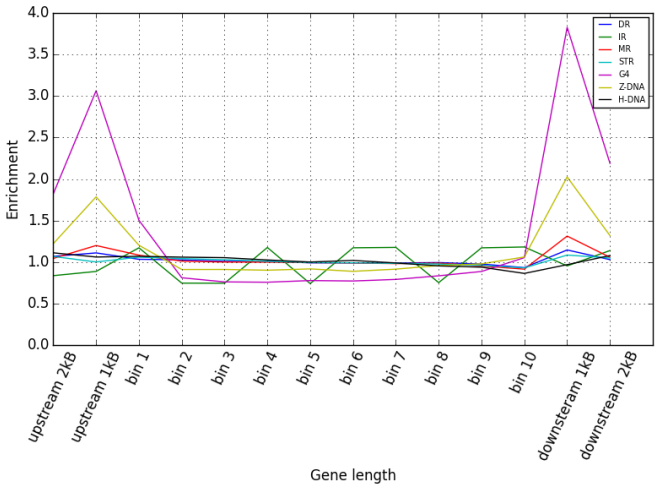
2.1.4. Distribution of non-B DNA motifs across the gene length.

To calculate the relative enrichment of each non-B DNA motif across the gene length, each gene was divided into ten equal sized bins. To those bins, two bins upstream and two bins downstream from the gene body each 1kB in length were added. The enrichment at each bin varied substantially between different non-B DNA motifs (Figure 2.3c). In addition, the distribution of non-B DNA motifs also varied across the gene body and were most enriched at promoters, 5'UTRs and 3'UTRs. In particular, G-quadruplexes displayed the highest enrichment levels at both ends of transcripts, followed by Z-DNA motifs. Therefore, these results suggest that multiple non-B DNA motifs are preferentially located upstream and downstream of the gene body, whereas they are less frequent found within the gene body.



a. Ensembl annotation

b. Segway annotation



c.

Figure 2.3: Non-B DNA motifs and genome partitions.

a. Distribution of non-B DNA motifs at gene regulatory regions as defined by the Ensembl regulatory features. b. Enrichment of occurrences of non-B DNA motifs at various chromatin states as defined by Segway annotation. c. Enrichment of non-B DNA motifs relative to their position in the gene body. The gene length was partiotioned in genomic bins in order to consider the disparities in size between different genes.

2.1.5. Positioning of non-B DNA motifs at functional sites at nucleotide resolution.

However, potential enrichment of non-B DNA motifs relative to specific functional sites in the genome and their positioning relative to them could not be accurately captured by calculating the total enrichment across entire regulatory regions which spanned hundreds or thousands of nucleotides. Therefore, 2kB window plots were generated centered at i) transcriptional start sites (TSSs), ii) transcriptional end sites (TESs), iii) coding sequence (CDS) start sites, iv) CDS end sites. For each of them the relative enrichment of each non-B DNA motif across the 2kB window was calculated at single nucleotide resolution.

Most of the non-B DNA motifs were enriched relative to the TSS, although their distribution peaks and enrichment levels varied (Figure 2.4a). The enrichment peak for Z-DNA motifs was at the TSS (~2.5-fold), whereas G-quadruplexes displayed two peaks upstream and downstream of the TSS (~1.7-fold). Similarly, inverted repeats and short tandem repeats displayed an enrichment peak in proximity to the TSS, ~1.35-fold and ~1.4-fold respectively. The enrichment relative to the TSS for multiple non-B DNA motifs indicated non-B DNA sequence preference at promoter regions and potentially the involvement of non-B DNA motifs in the regulation of gene expression.

For G-quadruplexes, the CpG-islands could be implicated in their enrichment at promoter regions. To address the contribution of CpG-islands which are often found overlapping TSSs, CpG-island annotations were retrieved from the UCSC genome browser. CpG-rich promoters were defined as the promoters for which the TSS overlapped with a CpG-island, whereas CpG-poor promoters were defined as the promoters whose TSS did not overlap CpG-islands. Promoters themselves were defined as +/-1kB from the 5' most TSS of each gene. The frequency of G-quadruplexes in each of the two categories was calculated. CpG-rich promoters had 1.047 G-quadruplexes per kB, while CpG-poor promoters had 0.545 G-quadruplexes per kB. By contrast, background G-quadruplex density across all Segway chromatin states was only 0.131 G-quadruplexes per kB.

Therefore, it was concluded that both CpG-rich and CpG-poor promoters are enriched for G-quadruplexes when compared to the background (~8-fold and ~4-fold respectively).

Next, the distribution of non-B DNA motifs relative to the TES was investigated. Here it was found that short tandem repeats were enriched (~2-fold), (Figure 2.4b). Inverted repeats (1.35-fold) and mirror repeats (1.3-fold) were also enriched, whereas G-quadruplexes were depleted. Although, on aggregate G-quadruplexes were depleted relative to the TES position, perhaps due to sequence constraints (as is shown in the next section, there is strand asymmetry and the pattern is more complex), their distribution downstream of the TES was explored in the previous section and it was found that G-quadruplexes are extremely enriched across the 3'UTR region, even more than in the promoter and 5'UTR regions, whereas they are depleted within the gene body (Figure 2.3c). In addition, the roles of G-quadruplexes at the 3'UTR have been recently investigated (Rouleau et al. 2017). Furthermore, in both CDS start and end sites, enrichment of short tandem repeats, Z-DNA and G-quadruplexes was evident. In particular, G-quadruplexes displayed two mirror peaks around the CDS start and end sites (Figure 2.4c-d). This is in support with experimental observations of G-quadruplex roles in translation initiation and termination (Halder et al. 2009), (Beaudoin and Perreault 2010).

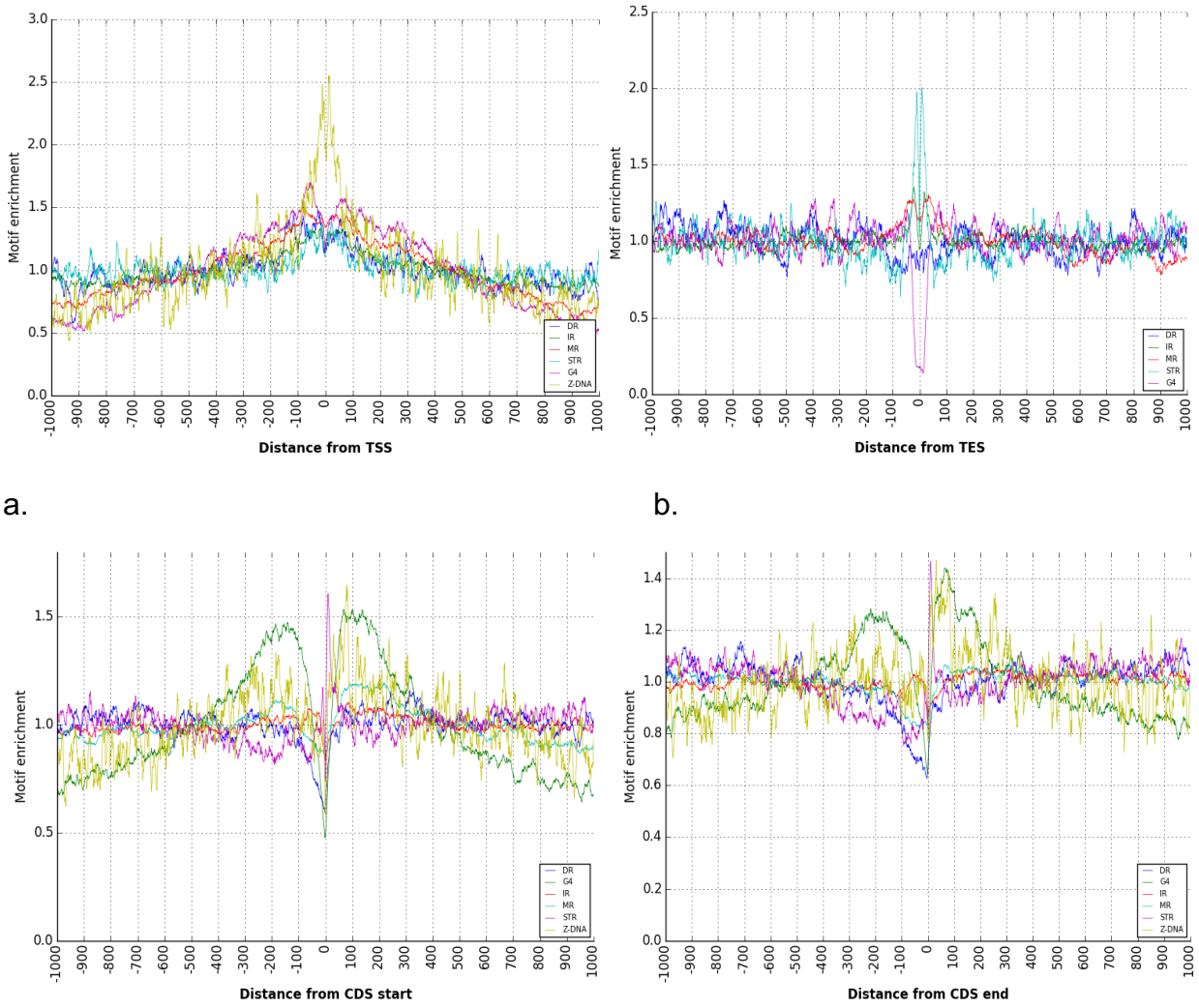


Figure 2.4: Non-B DNA motifs at transcription and translation start and end sites.

Enrichment of non-B DNA motifs around the a. TSS and the b. TES of coding genes. Enrichment of non-B DNA motifs around the c. CDS start and d. CDS end regions. H-DNA motifs were excluded from the generation of 2kB-window plots due to low numbers. Z-DNA motifs were excluded from panel b. due to low numbers.

G-runs and their enrichment at functional genomic sites. A script was developed to map G-runs across the human genome. Using those maps, a strong bias relative to the TSS was observed that was incremental from one to four G-runs, with an excess of G-runs in the non-template strand downstream of the TSS (Figure 2.5a). The bias was exaggerated with an increase in the number of G-runs, with four G-runs being the consensus G-quadruplex motif. At the TES there was non-template enrichment for G-runs, which was exaggerated for three and four G-runs (Figure 2.5b), which are more likely to fold in intermolecular and intramolecular G-quadruplex structures. Interestingly, the enrichment relative to the TES could not be observed by aggregating together G-quadruplex occurrences in the template and non-template strands, since the first was depleted and the second was enriched, therefore cancelling each other out. Similarly, at the CDS start, G-run enrichment was most pronounced for four G-runs at the non-template strand (Figure 2.5c), whereas at the CDS end such enrichment was not observed (Figure 2.5d). These results are very surprising, with most striking that of G-runs relative to the TES (Figure 2.5b) and implicated strand asymmetries in the distribution of G-quadruplexes at functional sites in the genome.

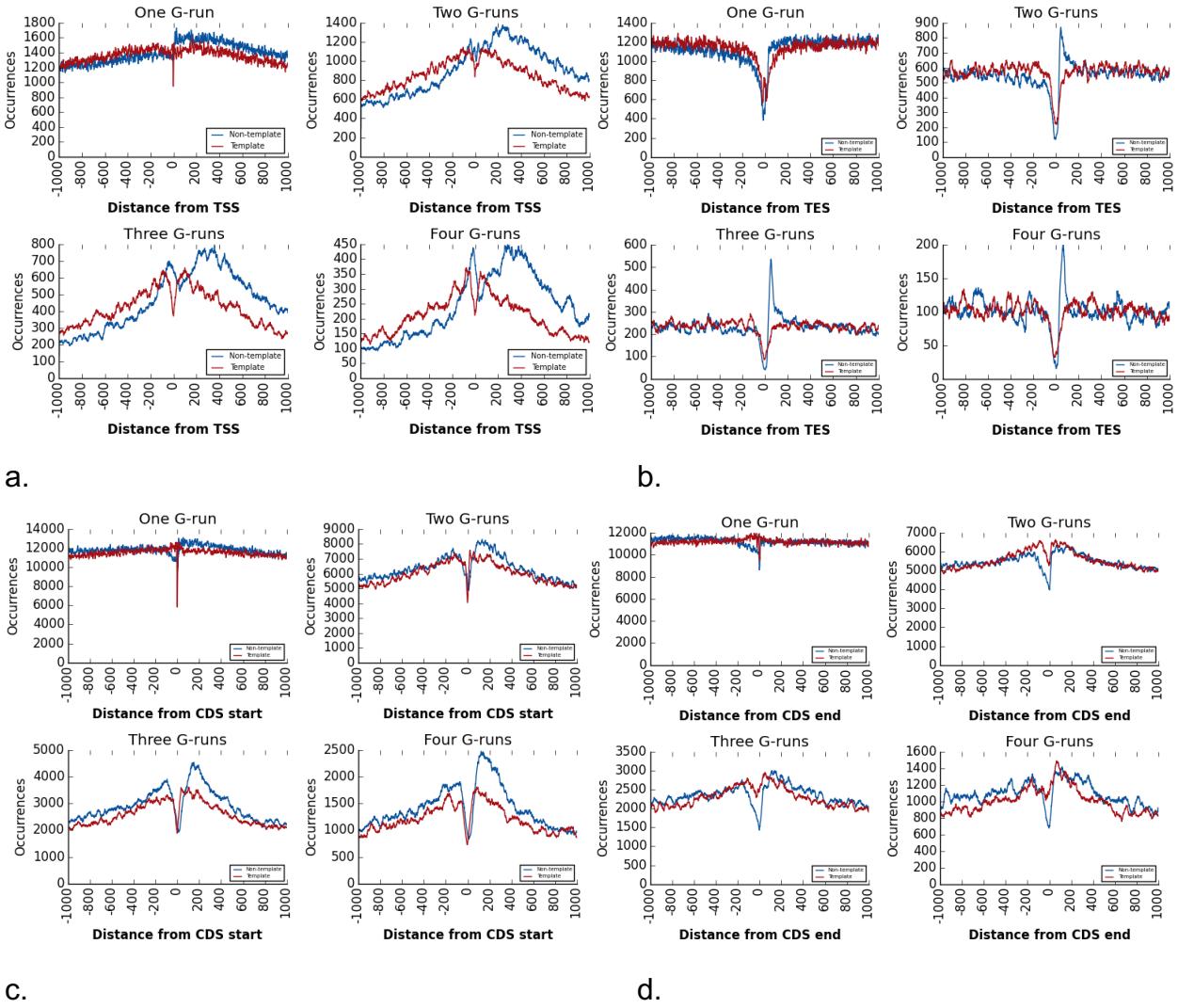


Figure 2.5: Transcriptional strand asymmetries associated with G-quadruplexes.

Template / Non-template strand asymmetry for one up to four consecutive G-runs interspersed with loops around a. the TSS (left) and b. the TES (right). Template / Non-template strand asymmetry of one to four consecutive G-runs interspersed by loops around c. the CDS start (left) and d. the CDS end (right). Four G-runs indicate the consensus motif for intramolecular G-quadruplex. The template strand is indicated in red and the non-template strand in blue.

2.2. Patterns of somatic mutations at non-B DNA motifs in multiple cancer genomes.

In total, 1,809 whole genome sequenced cancers were analysed, derived from 10 different tumour types. The analysis was focused on substitutions, small insertions and deletions (indels), and rearrangements. Among the 1,809 tumours, 560 of them were breast cancer (BRCA) tumours from (Nik-Zainal et al. 2016). Those available breast cancer genomes served as the largest WGS cohort of a single cancer type available at the time. It is a high-quality, manually-curated dataset and mutation calling was performed by the Wellcome Sanger Institute pipeline and a subset of mutations were experimentally validated. Therefore, this dataset served as the basis for the exploration of the roles of non-B DNA motifs in mutagenesis. The algorithms and methods were then validated in the other tumour types which were also run through the Wellcome Sanger Institute mutation calling pipeline (Table 2.1). The other cancer types were pancreatic (PACA), liver (LIRI), ovarian (OVCA), malignant lymphoma (MALY), pedriatic brain tumour (PBCA), renal carcinoma (RECA), prostate (PRAD), gastric (GACA) and esophageal (ESAD).

Table 2.1: Number of substitutions, indels and rearrangement breakpoints per tumour type.

Cancer Name	Samples	Substitutions	Indels	Rearrangement breakpoints
BRCA	560	3,479,651	371,993	131,068
LIRI	264	3,575,056	852,361	51,034
OVCA	72	732,189	141,296	39,078
ESAD	98	2,890,654	347,680	48,394
GACA	40	525,850	185,213	12,268
PBCA	239	299,241	231,874	13,120
PACA	242	1,881,336	625,803	48,404
RECA	74	584,144	123,180	1,972
MALY	100	1,242,356	203,051	10,752
PRAD	120	602,729	799,583	24,104

2.2.1. Mutational enrichment of non-B DNA motifs across the human genome.

Mutational density is uneven across the genome with genomic and epigenomic features contributing to differences in mutability (Polak et al. 2015). Non-B DNA motifs have been previously implicated in mutagenesis (Zhao et al. 2010), (Kamat et al. 2016). Nevertheless, systematic examination of cancer genomes to investigate mutational enrichment for each category of non-B DNA motifs has not been performed to date and it remains unclear what is the magnitude of their effect in mutagenesis.

To explore the mutational enrichment of non-B DNA motifs genome-wide, the genome was split in 500kB windows and analysed the likelihood of a mutation falling at a non-B DNA motif or within the vicinity of the window (Figure 2.6b). Firstly, genomic bins which did not have >50% of base pairs mappable including telomeric and centromeric regions were excluded, therefore avoiding sites with low mappability. Next, by measuring the coverage of each bin for each non-B DNA motif and the number of overlapping mutations the enrichment patterns were calculated. It was observed that both substitutions and indels are more likely to fall within non-B DNA motifs than in surrounding sequences (Figure 2.6a). Nevertheless, enrichment patterns could not be identified for rearrangements at most non-B DNA motifs, with the exception of inverted repeats (Figure 2.6a). Because of the lower number of rearrangements in comparison to substitutions and indels (between 1-2 orders of magnitude lower), (Table 2.1), potential roles of non-B DNA motifs in rearrangements cannot be excluded and there is evidence for putative roles of non-B DNA motifs in rearrangements from previous experimental studies (Bacolla et al. 2004), (Lu et al. 2015).

The enrichment levels were calculated for each non-B DNA motif and it was found that the differences were dependent on the non-B DNA motif category. For substitutions, the most enriched non-B DNA motifs were H-DNA, short tandem repeats and Z-DNA with 1.7-fold, 1.6-fold and 1.7-fold enrichment respectively (Figure 2.6a). Additionally, the other motifs were also found to be enriched, although the levels of enrichment were

smaller, G-quadruplex (1.2-fold), inverted repeat (1.1-fold), direct repeat (1.1-fold) and mirror repeat (1.1-fold). This analysis used their immediate surrounding sequence as background rate of mutagenesis, therefore correcting for potential epigenetic variants, genomic GC variation and replication timing differences. These results suggest that substitutions are enriched within non-B DNA motifs, which confer a higher probability of mutagenesis.

Indels showed even higher enrichment levels than substitutions across non-B DNA motifs. More specifically, Z-DNA (10.7-fold), H-DNA (6-fold), short tandem repeats (5.8-fold), mirror repeats (2.5-fold), direct repeats (2.3-fold) and G-quadruplexes (1.5-fold) were all found to be enriched for indel mutations (Figure 2.6a). These results are in accordance with the fact that indels across human cancers are often the result of insertion-deletion loops and replication slippage, which could occur more frequently at non-B DNA motifs. Finally, for rearrangements enrichment was observed within inverted repeats in breast cancer (1.2-fold) (Fig. 2.6a). This finding is in support with experimental observations in yeast and mammalian *in vitro* and *in vivo* studies (Lu et al. 2015). Overall, these results indicate enrichment of mutagenesis at non-B DNA motifs at large genomic bins for multiple types of mutations. Also, to provide further evidence additional analysis was performed by exploring mutational patterns at non-B DNA motifs at a nucleotide resolution.

2.2.2. Mutational enrichment at non-B DNA motifs at nucleotide resolution.

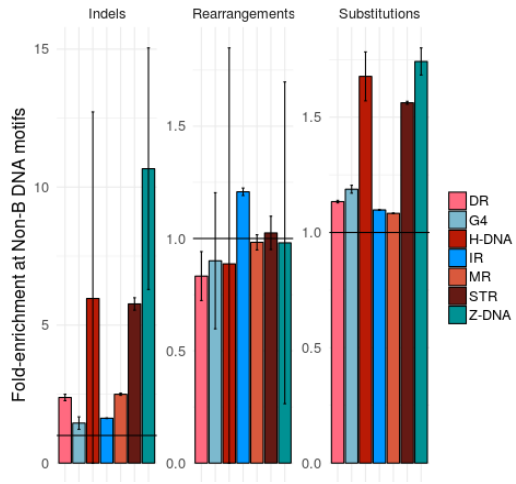
To further investigate the distribution patterns of non-B DNA motifs at mutation sites, 2kB window plots were generated centered at each type of somatic mutation. Next the distribution of each non-B DNA motif relative to the mutation site was investigated. Similar to the results described earlier, it was found that non-B DNA motifs are enriched for substitution mutations and the enrichment level varies by non-B DNA motif (Figure 2.6c). In particular, the enrichment of non-B DNA motifs is directly at the site of mutagenesis,

therefore suggesting that the previous results (Figure 2.6a) were not due to confounding factors.

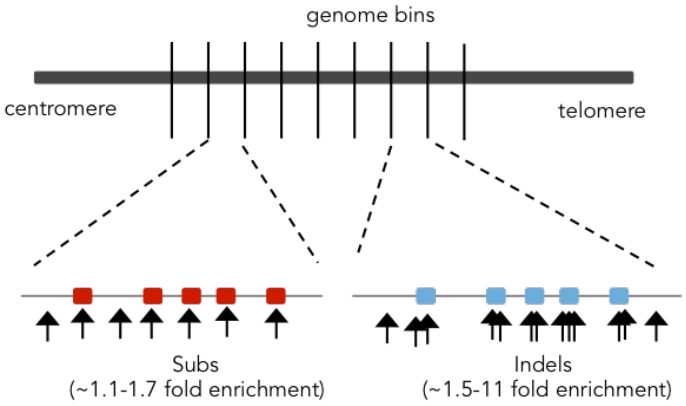
Mutational enrichment at non-B DNA motifs at nucleotide resolution, controlling for trinucleotide context. For substitutions, the trinucleotide context of mutagenesis contributes to the observed mutational patterns in cancer genomes. As a result, the trinucleotide context could also contribute to the observed enrichment at non-B DNA motifs. To estimate its effects, control mutations were simulated controlling both for: i) the trinucleotide content of mutations, and ii) the genomic localisation, across each tumour type independently. Next, the observed enrichment was calculated in comparison to the expected based on the simulations. It was found that the patterns of enrichment at non-B DNA motifs were largely unchanged after controlling for the trinucleotide context (Figure 2.6d). Results are shown independently for each motif across tumour types in (Georgakopoulos-Soares et al. 2018) with almost unchanged enrichment patterns, with the exception of Z-DNA, for which the enrichment levels decreased; however, the Z-DNA motif remained enriched relative to the controls. Therefore, these results suggest that the observed enrichment at non-B DNA motifs is not the result of the trinucleotide context of the mutation.

Similarly, indel mutations are enriched across all non-B DNA motifs. The nucleotide resolution plots indicate that mutations precisely fall at non-B DNA motifs with higher likelihood, which further suggests their direct implication in cancer mutagenesis. The enrichment level is variable between non-B DNA motifs and more pronounced than that observed at substitutions (Figure 2.6c-e). Across most non-B DNA motifs, the enrichment is centered at the mutation site. However, for G-quadruplexes a central peak of enrichment was observed at the indel site and two additional symmetric peaks were observed ~150bp away, across multiple cancer types (Figure 2.6f). Further analysis indicated a link between presence of G-quadruplexes and nucleosome-free regions (Figure 2.6g). Nucleosome occupancy has been previously shown to be a major determinant of indel mutability (Morganella et al. 2016); here it is shown that G-quadruplexes are associated with nucleosome positioning and indel mutagenesis

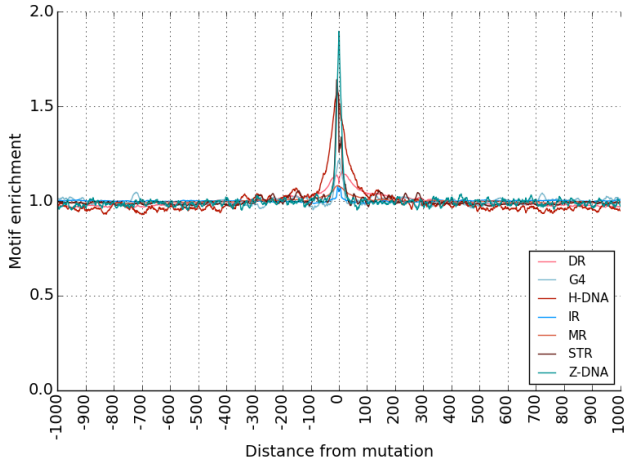
(Figure 2.6f-g). Therefore, the association between G-quadruplex positioning and nucleosome formation shapes the indel profile of multiple cancer genomes.


Indels
Rearrangements
Substitutions
Fold-enrichment at Non-B DNA motifs
DR
G4
H-DNA
IR
MR
STR
Z-DNA

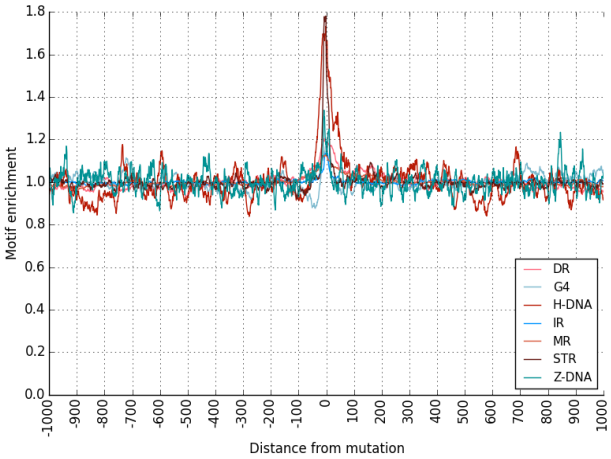
a.


genome bins
centromere
telomere
Subs
(~1.1-1.7 fold enrichment)
Indels
(~1.5-11 fold enrichment)

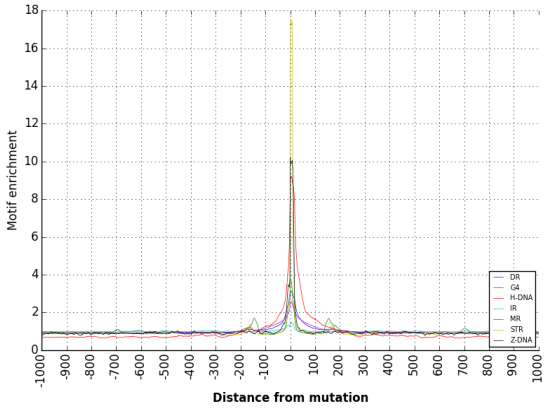
b.


Motif enrichment
Distance from mutation
DR
G4
H-DNA
IR
MR
STR
Z-DNA

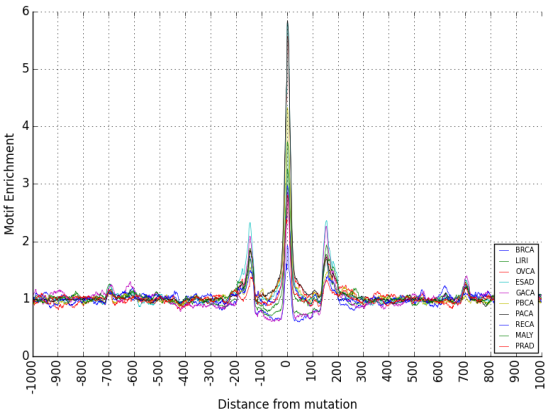
c.


Motif enrichment
Distance from mutation
DR
G4
H-DNA
IR
MR
STR
Z-DNA

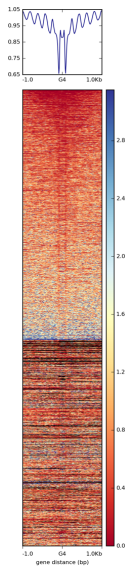
d.


Motif enrichment
Distance from mutation

e.


Motif Enrichment
Distance from mutation
BRCA
LIRI
OVCA
ESAD
GACA
PBCA
PACA
RECA
MALY
PRAD

f.



g.

Figure 2.6: Non-B DNA motifs are enriched for substitutions and indels across cancer types.

a. Enrichment of non-B DNA motifs at genomic bins for substitutions, indels and rearrangements. b. Depiction of enrichment per genomic bin, for results in (a), demonstrating how mutations are enriched at non-B motifs. Red and blue boxes represent non-B motifs. c. Mean enrichment at each non-B DNA motif for substitutions across ten cancer types. d. Mean enrichment at each non-B DNA motif for substitutions across ten cancer types, controlling for the trinucleotide context biases of substitution mutations. e. Mean enrichment at each non-B DNA motif for indels across ten cancer types. f. G-quadruplex distribution relative to indel sites across the different tumour types. g. The relationship between G-quadruplexes and nucleosome positioning is shown, using MNase data from the K562 cell line.

2.3. Discussion.

Non-B DNA motifs are found in higher frequencies than expected by chance in the human genome (Schroth and Ho 1995), (Cox and Mirkin 1997). Importantly, they are not evenly distributed along the genome; they are preferentially located in certain genomic sites. Because of their frequent occurrences at some functional elements, they could have strong and direct implications in the regulation of gene expression. Indeed, disruption of non-B DNA motifs in promoters can dramatically alter expression levels of the downstream gene (Siddiqui-Jain et al. 2002), but their wider implications perhaps remain underappreciated given their frequent occurrences and relative positioning at regulatory sites. For G-quadruplexes, an in-depth analysis was performed focusing on strand asymmetry and G-run enrichments. Strong strand asymmetry was observed between template and non-template strands, which is most extreme around the TES, in which G-quadruplexes are ~2-fold enriched at the non-template strand, but de-enriched at the template strand (Figure 2.3c, Figure 2.4b, Figure 2.5b). Future experiments would be needed to further explore the roles of non-B DNA motifs at regulatory regions and investigate the implications of the observed asymmetries.

Mutational density was higher at non-B DNA motifs than in surrounding regions across cancer types and varied by non-B DNA motif (Georgakopoulos-Soares et al. 2018), (Figure 2.6a-b). Additionally, their enrichment levels were pinpointed with window plots,

indicating their direct implication in increased mutability both in substitutions and indels (Figure 2.6c-g). Moreover, in the case of G-quadruplexes at indels, a surprising enrichment of G-quadruplexes was observed approximately 150bp away from the site of the mutation on either side, in addition to the enrichment peak directly overlapping indels (Figure 2.6f). This was shown to be due to the positioning of G-quadruplexes relative to nucleosomes (Figure 2.6g). It remains unknown which processes and potential mechanisms could underlie the enrichment of indels at G-quadruplexes in relationship to nucleosome positioning. Moreover, the interplay between G-quadruplexes and nucleosome is indicative of complex mechanisms that can underlie mutagenesis.

For indels, the enrichment of non-B DNA motifs was more dramatic than that of substitutions (Figure 2.6c-f), which is in accordance with previous experimental studies indicating the role of non-B DNA motifs in indel mutagenesis (Collins 1981), (Freund et al. 1989), (Lu et al. 2015), (Lobachev et al. 1998). Insertion-deletion loops form frequently at repeat sites; which likely explains the strong enrichment at tandem repeats. Furthermore, in MMR deficient tumours the enrichment levels at tandem repeats were exacerbated, indicative of the role of mismatch repair in repairing DNA damage at these structures (Richman et al. 2015). Finally, it remains unclear if the disruption of non-B DNA motifs by substitution and indel mutations could have implications in cancer development.

CHAPTER THREE

3. Non-B DNA motifs are determinants of mutability in cancer genomes.

In this chapter a genome-wide model of mutability is constructed using epigenetic modifications, replication timing and non-B DNA motifs as inputted features to predict mutational patterns across the genome for multiple cancer types. Furthermore, it is shown that there are differences in the mutability within non-B DNA motif sub-components across multiple non-B DNA motif categories and that more exposed regions are more mutable for substitution mutations. Sequence characteristics, such as spacer and arm length in direct, mirror and inverted repeats contribute to observed differences in mutability, with spacer sequences showing an excess of mutations relative to the arms. Similarly, in G-quadruplexes a higher mutational density is observed at the G-runs relative to the looping regions. Finally, it is shown that non-B DNA motifs contribute to locally elevated rates of mutability resulting in an enrichment for recurrent mutagenesis across cancer patients.

3.1. Introduction: Epigenetic and non-B DNA motif influences on mutability.

The distribution of somatic mutations along cancer genomes is largely heterogeneous. The mutational density varies substantially between different genomic locations, with a range of more than five-fold (Lawrence et al. 2014). At the base-pair scale, the trinucleotide context at the mutational site has been strongly associated with differences in mutability, which reflect specific mutational processes (Nik-Zainal et al. 2012a), (Alexandrov et al. 2013). At the mega-base scale, epigenetic modifications and replication

timing have been previously shown to be major determinants of mutability (Polak et al. 2015). In particular, heterochromatin and late replicating regions contain an excess of substitutions, whereas active chromatin, early replicating sites and transcribed regions display lower mutational densities. However, occupied transcription factor binding sites at open chromatin regions have a higher mutational rate than the surrounding regions, likely reflecting inaccessibility of those sites to repair enzymes (Sabarinathan et al. 2016). As a result, the mutational landscape of cancer genomes is influenced by numerous features, with multiple and often complex relationships.

Here the relationship between non-B DNA motifs and mutation rates was investigated across multiple cancer types. In addition, their potential exploitation to predict genome-wide mutation rates was examined and compared to that of epigenetic modifications and replication timing. Following the global analysis, mutational patterns at non-B DNA motifs were analysed at the nucleotide level. More specifically, the relationship between specific characteristics of non-B DNA motifs and the likelihood of mutagenesis was also examined. Finally, it was shown that non-B DNA motifs are recurrently mutated across cancer types, therefore obfuscating the interpretation of recurrently mutated sites across the genome.

3.2. Analysis of epigenetic and non-B DNA motif influences on mutability across cancer genomes.

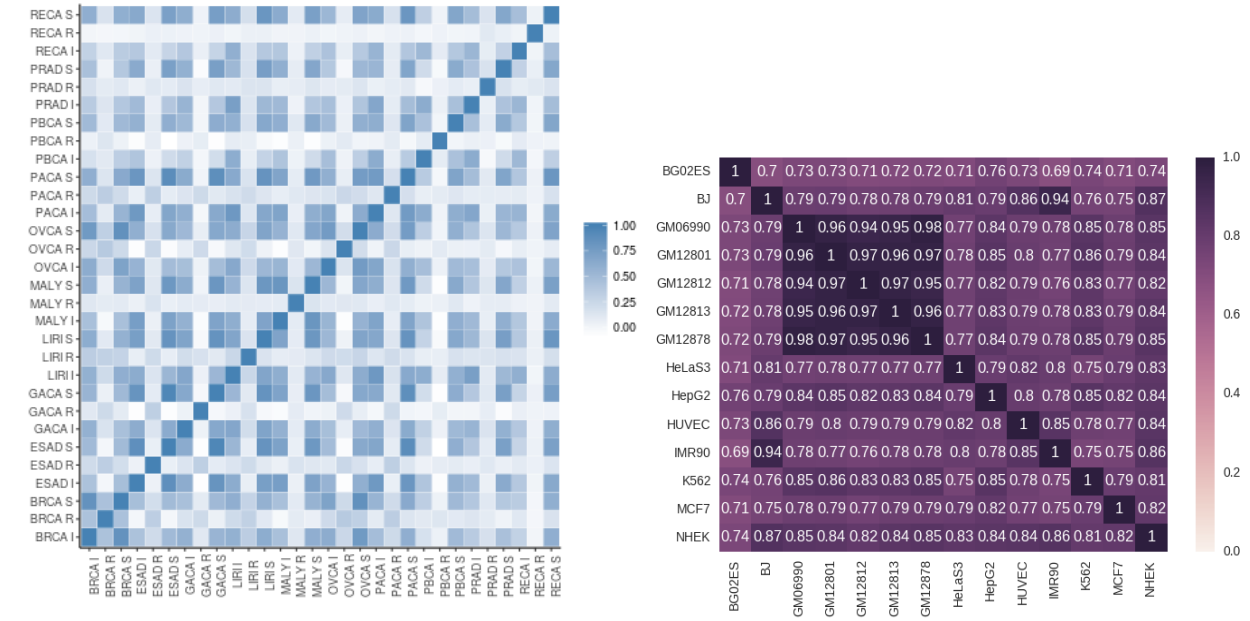
The human genome was binned in 500kB windows. Centromeric, telomeric and bins with consistently low mappability were excluded. The number of substitutions, indels and rearrangements was calculated for each bin and compared between cancer types using simple Pearson correlations (Figure 3.1a). Next, the association between mutational density and a multitude of genomic and epigenomic features was investigated including: i) seven non-B DNA motifs, ii) seven epigenetic modifications, and iii) replication timing. The number of occurrences of each non-B DNA motif category in each of the genomic

bins was measured. In addition, the nucleotide coverage, defined as the proportion of overlapping nucleotides, at each genomic bin for each epigenetic modification, derived from a cell of origin for each tumour type, was calculated. These included H3K9me3 which corresponded to heterochromatin regions, H3K4me1 and H3K4me3, which marked different regulatory elements such as enhancers and promoters, H3K36me3, which is present in transcribed regions, DNase which characterises accessible, open chromatin regions and H3K27ac which corresponds to active genomic elements. Finally, to investigate the relationship between mutational density and replication timing, Repli-seq data were mapped to the genomic bins. For replication timing data, the corresponding cell of origin for each tumour type was not always available. However, the Pearson correlation between any two Repli-seq datasets was >0.69, which indicated a high degree of similarity between different cell lines (Figure 3.1b). With this plethora of described data, it was explored how each of those features correlated with each of the mutation types.

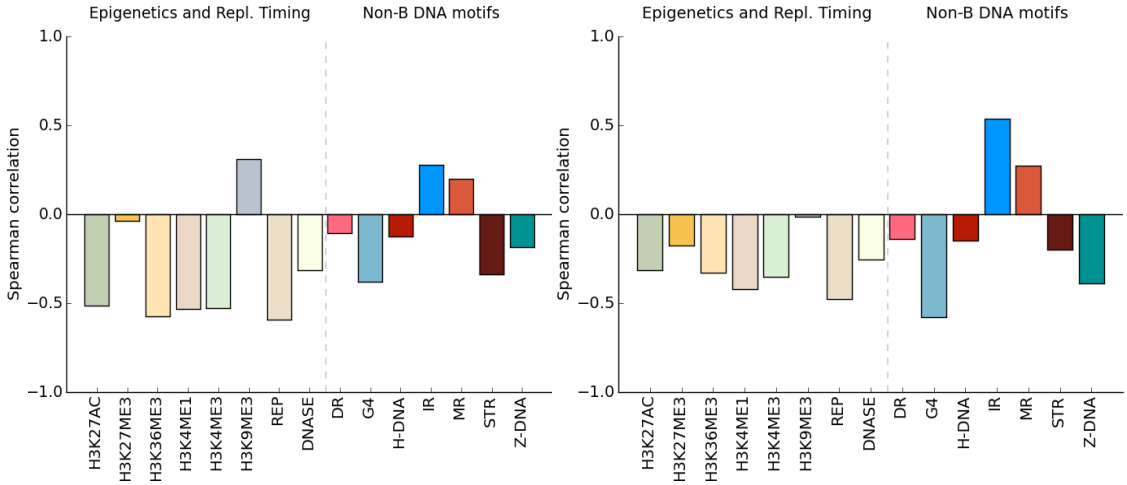
First, simple linear relationships between non-B DNA motifs, epigenetic modifications and replication timing and mutability were investigated. Consistent with previous studies (Schuster-Böckler and Lehner 2012), (Polak et al. 2015), it was found that heterochromatin marks, denoted by H3K9me3 and late replicating domains are correlated with increased mutational density, r=0.31 and r=0.59 respectively. In contrast to that, open chromatin regions identified by DNase, active regulatory elements and transcribed regions were negatively correlated with mutability, r=-0.31, r=-0.52 and r=-0.57 respectively (Figure 3.1c). The correlations for non-B DNA motifs were: inverted repeats (r=0.28), short tandem repeats (r=-0.33), G-quadruplexes (r=-0.38), mirror repeats (r= 0.20) and Z-DNA (r=-0.19) (Figure 3.1c). The strong correlations observed between non-B DNA motifs and mutability indicated that similar to epigenetic and replication timing domains, non-B DNA motifs could be used as predictive features of genome-wide mutational density. However, an additional benefit of using non-B motifs is that they can be derived from the primary reference genome sequence, in contrast to epigenetic marks and replication timing. Similar to substitutions, strong correlations between indels and genomic and epigenomic features were found (Figure 3.1d). However, for rearrangements strong correlations could not be observed for any of the epigenetic

marks, replication timing or non-B DNA motifs. These results were consistent across all cancer types although only substitution and indel results for breast cancers are displayed (Figure 3.1.c-d), while the data for the other cancer types and for rearrangements can be found in (Georgakopoulos-Soares et al. 2018). The weak correlations for rearrangements could be due to the lower number of rearrangements relative to indels and substitutions (Table 2.1), resulting in higher uncertainty levels or due to additional processes being influential.

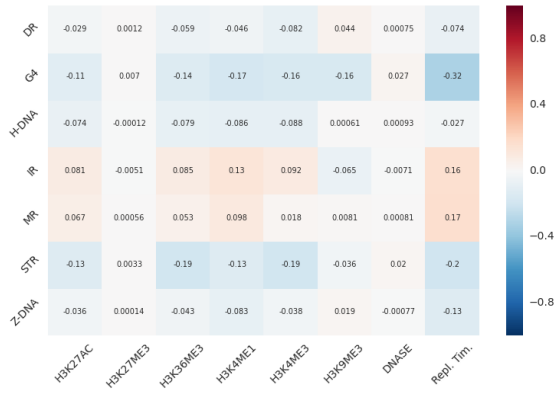
Partial correlations were implemented to investigate the contribution of a third feature in the association between two variables. Partial correlation analysis was implemented and revealed that the association between somatic mutations and non-B motifs remained even after controlling for each of the epigenetic marks and for replication timing, even though the most associated covariate was replication timing (Figure 3.1e), raising the possibility that non-B motifs could contribute to a predictive model of mutagenesis.



a. b.



c. Substitutions d. Indels



e.

Figure 3.1: Association between somatic mutations and non-B DNA motifs, epigenomic features and replication timing.

a. Pearson correlation between the number of substitutions (S), indels (I) and rearrangements (R) found in non-overlapping 500kB bins across the ten tumour types. b. Correlations between replication timing domains of different cell lines (Pearson, 500kb genome windows). Correlations between the number of non-B DNA motifs, epigenetic features and replication timing with: c. the number of substitutions for breast cancer (Spearman correlation), d. the number of indels for breast cancer (Spearman correlation). e. Results following partial correlation analyses. Remaining correlations (Pearson - partial) for each non-B motif (y-axis) when controlling for epigenetic features and replication timing (x-axis). Across the panels, IRs represent inverted repeats, DRs direct repeats, MRs mirror repeats, G4s G-quadruplexes and STRs short tandem repeats.

To further these claims, a principal component analysis was performed and revealed that epigenetic features and non-B DNA motifs are separated by the first two principal components (Figure 3.2a). This associates them with mutability in distinct ways and is further supported by the partial correlation analysis. Crude correlations indicated that genomic and epigenomic features could be used predictively. Furthermore, non-B DNA motifs can be inferred from the primary DNA sequence alone, therefore generating a cost-effective method to improve mutability predictions. Therefore, a genome-wide model of mutability was built to investigate the predictive ability of: i) non-B DNA motifs, ii) epigenetic features and replication time domains, iii) the combination of non-B DNA motifs, epigenetic features and replication time domains. Two distinct algorithms were developed. The first constructed model implemented linear regression and captured basic linear relationships. The second model implemented random forest regression and was more accurate than the linear model (Georgakopoulos-Soares et al. 2018), because it could capture more complex relationships. More specifically, the random forest is a popular machine learning algorithmic model of ensemble learning, which is based on the construction of multiple decision trees, which are then combined towards the final model generation for prediction or classification problems. The results obtained from the random forest model are in accordance with previous analysis indicating the superiority of random forest regression when compared to linear models (Polak et al. 2015).

For breast cancers, non-B DNA motifs explained 37% of the observed variance in substitution mutational density, but when the model also incorporated epigenetic modifications and replication timing, the variance explained reached 52%, performing better than either epigenetics and replication timing or non-B DNA motifs alone (Figure 3.2b). No improvement was observed when instead of non-B DNA motifs, GC-content or gene information was added in the predictive models. In addition to the strong correlations, H3K9me3 and replication timing were found to be highly informative features in predicting mutational density across tumour types (Figure 3.2c), (Georgakopoulos-Soares et al. 2018). Moreover, it was found that inverted repeats and G-quadruplexes were strong predictors, while multiple other non-B DNA motifs contributed to prediction accuracy (Figure 3.2d).

For the prediction of indel mutability across the genome using the random forest regression model, the non-B DNA motif model performed similarly to the epigenetics and replication timing model across tumour types, while combined the predictive ability was substantially increased (Georgakopoulos-Soares et al. 2018). Finally, for rearrangements both the predictive model of non-B DNA motifs and that of epigenetics and replication timing performed poorly (Georgakopoulos-Soares et al. 2018), raising the possibility that either other features that were not included are important predictors or that the lower number of rearrangements in comparison to that of indels and substitutions was the underlying reason. Therefore, it was concluded that future models of mutability for substitutions and indels should include non-B DNA motifs as additional features when predicting mutability across the human genome.

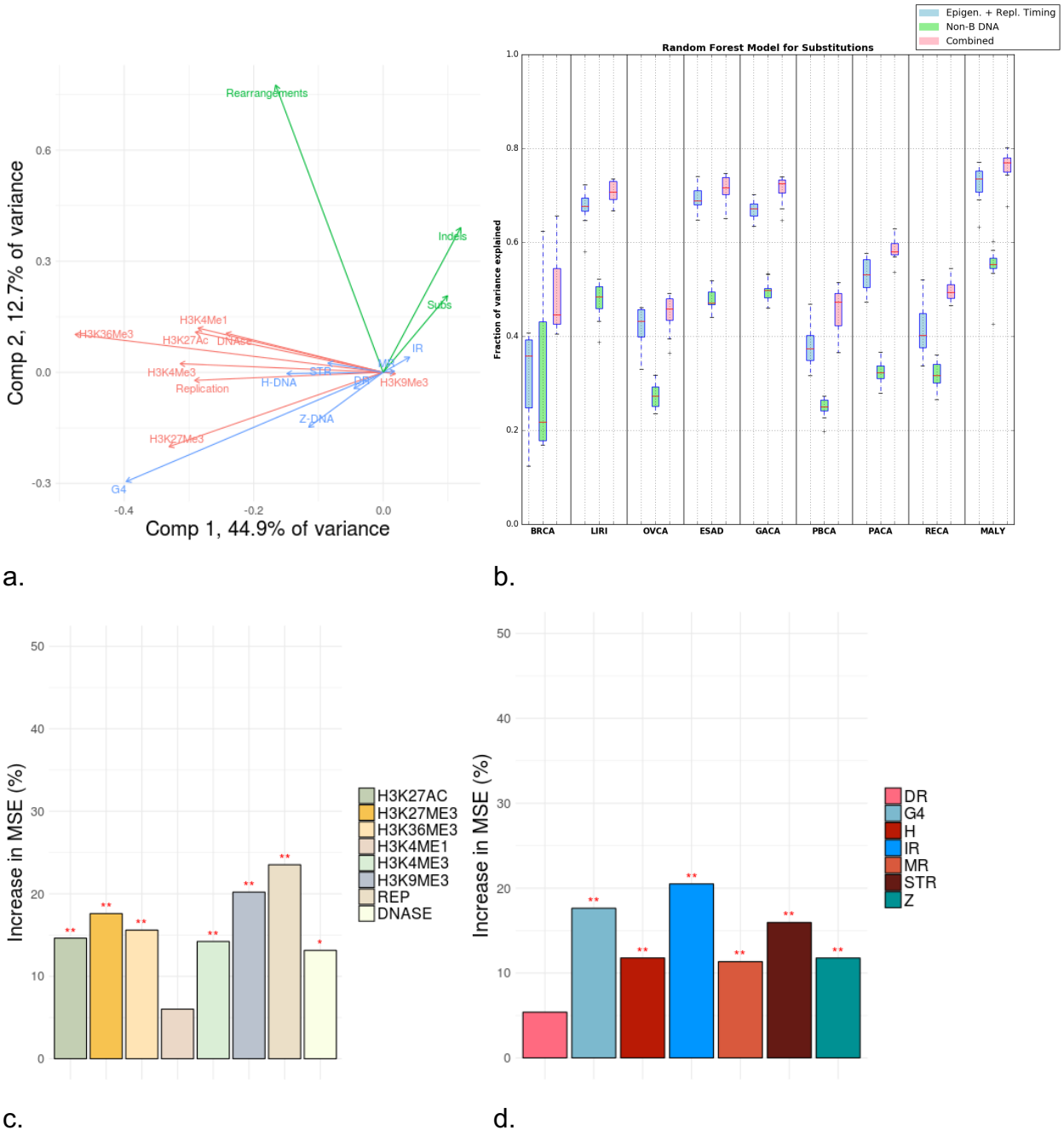


Figure 3.2: Non-B DNA motifs predict somatic mutability in cancer genomes.

a. PCA Analysis. The first two principal components separate mutations (green), non-B DNA motifs (blue) and epigenetics and replication timing domains (red). B. Fraction of variance explained for predicting the number of mutations in 500 kb bins with random forest regression using non-B DNA motifs and epigenetic features/replication timing as predictors for multiple tumor types (BRCA = breast cancer, LIRI=liver cancer, OVCA =ovarian cancer, ESAD = esophageal adenocarcinoma, GACA = gastric cancer, PBCA = pedriatic brain cancer, PACA = pancreatic cancer, RECA = renal cell carcinoma, MALY = malignant lymphoma). Error bars represent standard error from 10-fold cross-validation. (c, d) Importance of the different predictors

for the random forest regression. The y-axis shows the increase in mean square error (MSE) when the variable is excluded. Bars with * have an FDR<.05 and ** have FDR<.01 as determined by a permutation test. Across the panels, IRs represent inverted repeats, DRs direct repeats, MRs mirror repeats, G4s G-quadruplexes, STRs short tandem repeats, Z Z-DNA and H H-DNA.

3.3. Sequence characteristics of non-B DNA motifs and mutability.

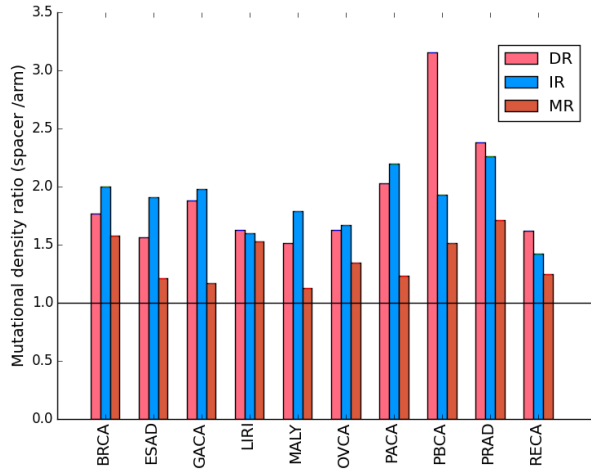
Mutations were found to be enriched at non-B DNA motifs across tumour types in chapter two; here non-B DNA motif differences in mutational enrichment were investigated depending on sequence characteristics and properties, specific to each non-B DNA motif. In addition, previous experiments have indicated that the sequence characteristics of each non-B DNA motif determine the likelihood of secondary structure formation and its associated thermodynamic stability (Nag and Petes 1991), (Rentzeperis et al. 1993), (Tippana et al. 2014). More specifically, a multitude of previous experiments have also indicated that hairpin formation is influenced by the spacer and arm lengths and their nucleotide composition (Sinden et al. 1991), (Lobachev et al.1998), while G-quadruplex stability is dependent on its loop length, with smaller loop lengths favouring G-quadruplex formation (Tippana et al. 2014), (Piazza et al. 2015).

Non-B DNA motif subcomponents could display differences in their mutational densities across cancer genomes, depending on how exposed they are and how often they are found single stranded (Figure 2.1). Indeed, previous studies have provided evidence for the hypermutability of spacer sequences at inverted repeats (Weinhold et al. 2014), (Nik-Zainal et al. 2016). The suggested mechanism implicates the formation of hairpin structures at inverted repeats, of which the spacer sequence is single stranded and exposed therefore harbouring an excess of mutations in comparison to the double-stranded arms. Similarly, during G-quadruplex formation, the loop regions are single stranded and more exposed and could therefore also harbour an excess of mutations. Here, mutational enrichment was shown to be influenced by the sequence characteristics

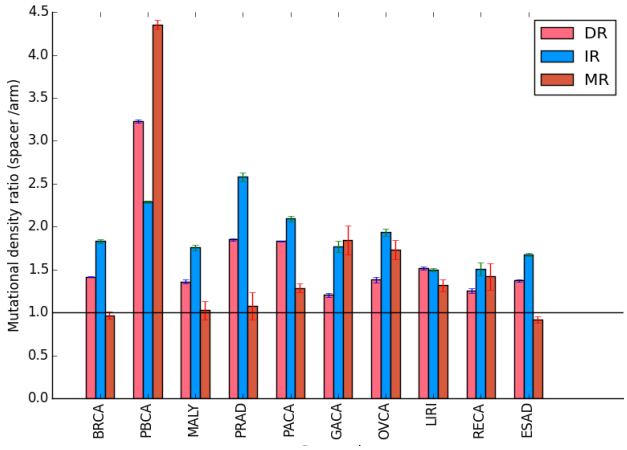
of non-B DNA motifs and be non-homogeneous across the sub-components of each non-B DNA motif.

Differences in the spacer to arm mutational enrichment for inverted repeats, direct repeats and mirror repeats were observed across cancer types (Figure 3.3a). It was found that spacer sequences are more enriched for substitutions than arm sequences (1.8-fold for direct repeats, 2-fold for mirror repeats and 1.7-fold for inverted repeats) (Fig. 3.3a). These differences could not be explained by the trinucleotide context of substitutions. To support this claim, simulated mutations were generated correcting for the trinucleotide content and genomic proximity to the actual mutations. Even after correcting for the trinucleotide context using the simulated mutations it was observed that spacers had a higher mutational enrichment than the arms (Figure 3.3b). As a result, it was concluded that subcomponents of direct, inverted and mirror repeats are preferential targets of mutability.

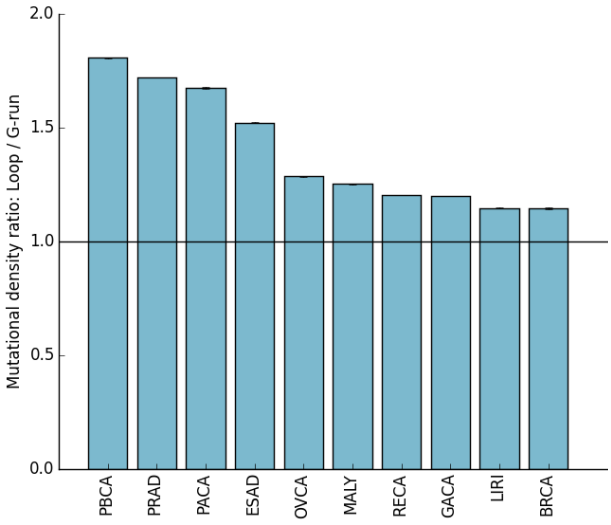
G-quadruplexes are composed of G-runs interspersed by loop elements, of which the second are single stranded and exposed during structure formation. Previous experiments show that smaller loops lead to higher likelihood of G-quadruplex formation and higher thermodynamic stability (Tippana et al. 2014), (Piazza et al. 2015) and at the same time it was hypothesized they are more prone to mutagenesis. It was found that loop regions are more enriched for mutations than the G-runs (Figure 3.3.c) and the trinucleotide context of substitution mutations does not explain differences in G-run and loop mutability (Figure 3.3d). Additionally, the subset of G-quadruplexes that have smaller looping regions have a higher mutational density than G-quadruplexes with longer average loop length (Figure 3.3e).


Mutational density ratio (spacer /arm)
DR
IR
MR
BRCA
ESAD
GACA
LIRI
MALY
OVCA
PACA
PBCA
PRAD
RECA

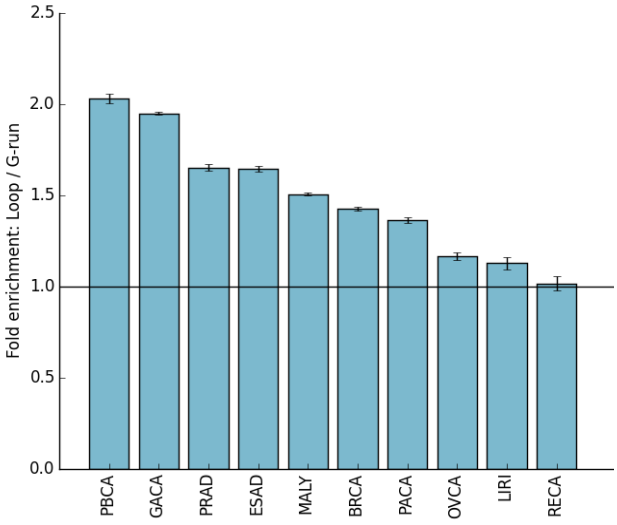
a.


Mutational density ratio (spacer /arm)
DR
IR
MR
BRCA
PBCA
MALY
PRAD
PACA
GACA
OVCA
LIRI
RECA
ESAD

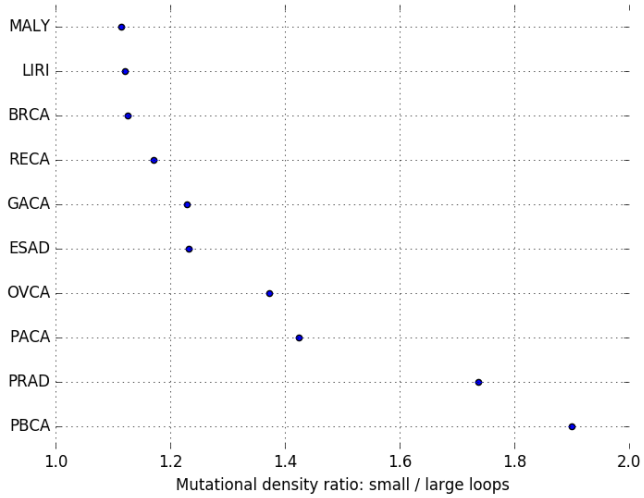
b.


Mutational density ratio: Loop / G-run
PBCA
PRAD
PACA
ESAD
OVCA
MALY
RECA
GACA
LIRI
BRCA

c.


Fold enrichment: Loop / G-run
PBCA
GACA
PRAD
ESAD
MALY
BRCA
PACA
OVCA
LIRI
RECA

d.


MALY
LIRI
BRCA
RECA
GACA
ESAD
OVCA
PACA
PRAD
PBCA
Mutational density ratio: small / large loops

e.

Figure 3.3: Increased mutability is domain-specific for particular non-B DNA motifs.

a. Mutational density in spacers compared to arms for direct repeats, inverted repeats and mirror repeats. b. Mutational density in spacers compared to arms for direct repeats (DRs), inverted repeats (IRs) and mirror repeats (MRs), corrected for trinucleotide context of substitutions. Error bars represent standard error from bootstrapping. c. Enrichment of mutational density in loops over G-runs across ten cancer types. Error bars represent standard deviation from bootstrapping with replacement (n=10,000). d. Enrichment of mutation density in loops: G-runs across ten cancer types corrected for trinucleotide content. Error bars represent standard deviation from bootstrapping with replacement (n=10,000). e. Enrichment of mutation density at G-quadruplexes for small loop sizes (less and equal to 3nt) relative to large loop sizes (more than 3nt) across ten cancer types. Error bars represent standard error from bootstrapping with replacement (n=10,000). Mann-Whitney *U* test was performed for each cancer type (p-value < 0.001 across all tumour types).

Furthermore, it was investigated how the spacer to arm mutational enrichment is influenced by the spacer and arm length of inverted, direct and mirror repeats. Heatmaps of mutational enrichment were constructed which indicated that for inverted repeats there was a cluster of spacer to arm mutational enrichment for spacer sequences of 1-3bps and arm lengths of 10-14bps (Figure 3.4a, d, e). For direct repeats, shorter spacer lengths were associated with increased mutability at spacers versus arms (Figure 3.4b,d). In accordance to this, longer spacer sequences impede the formation of slipped structures (Pearson et al. 1998). Mirror repeats were found to have less pronounced spacer to arm enrichment differences for spacer and arm length changes (Figure 3.4c-e), (Figure 3.3a-b). Nevertheless, the subset of mirror repeats that have high AG content (>90%) and can fold in H-DNA structures (Figure 2.1) were found to be highly mutagenic (Figure 2.4a). These results implicate the sub-components of non-B DNA motifs in domain-specific differences of mutagenesis. In addition, physical characteristics of non-B DNA motifs are implicated in their respective differences of mutability.

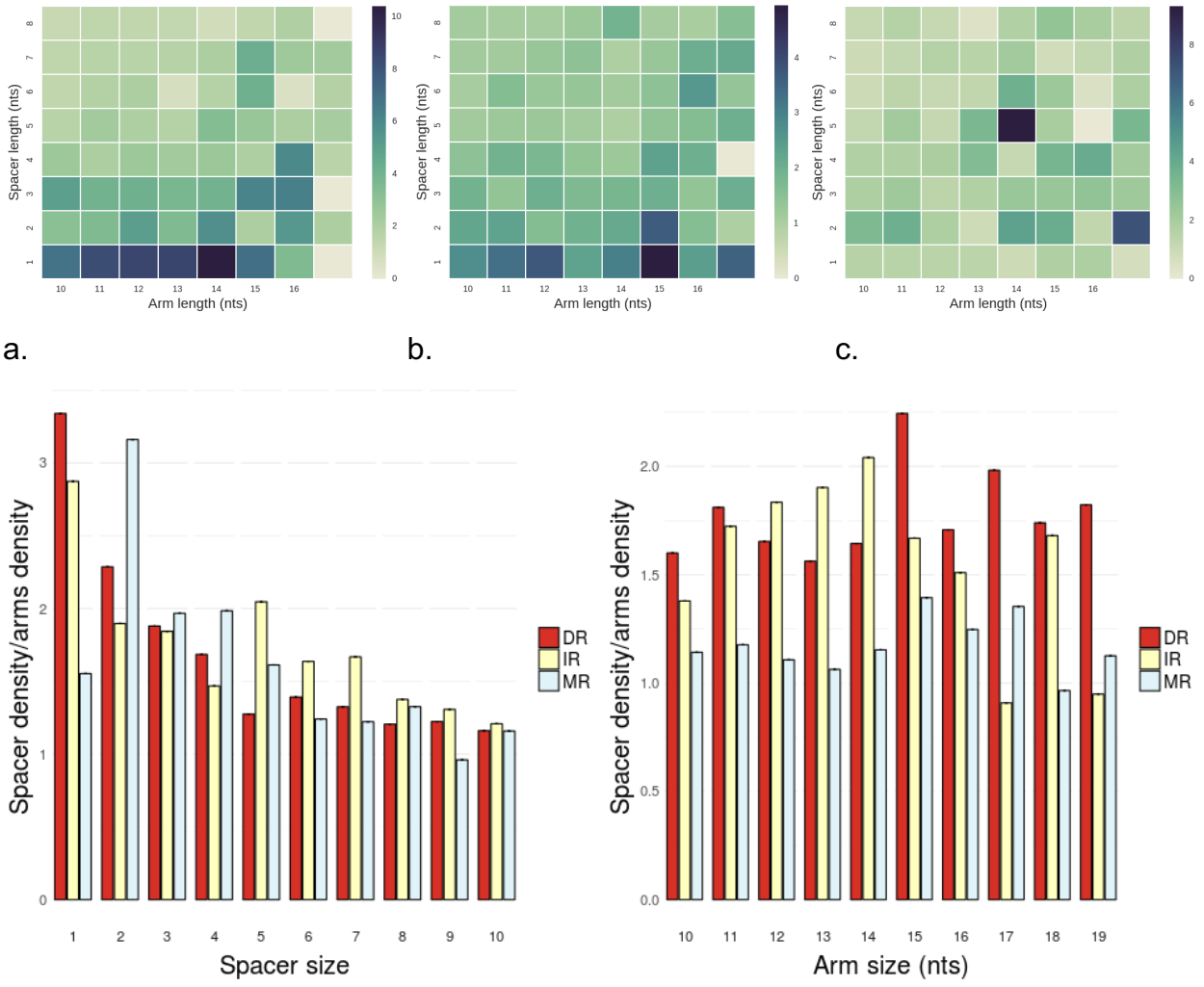


Figure 3.4: Mutability is dependent on the sequence characteristics of non-B DNA motifs and varies between their sub-components.

Heat map showing relative ratio of mutational density of spacers over arms for breast cancer at a. inverted repeats (IRs), b. direct repeats (DRs) and c. mirror repeats (MRs). d. Mutational density at spacer versus arm by spacer size for breast cancers. e. Mutational density at spacer versus arm by arm size for breast cancers.

3.4. Recurrency of mutagenesis at non-B DNA motifs.

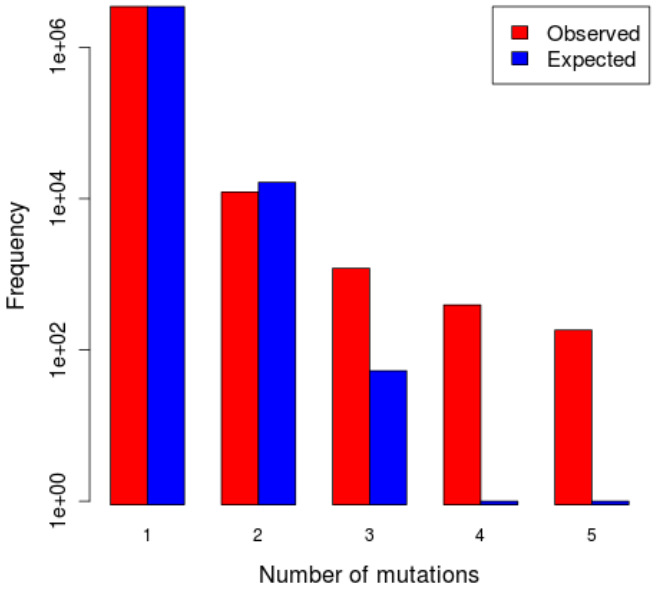
Identification of driver mutations in non-coding regions of the genome remains challenging. There are multiple reasons for the difficulties associated with finding such driver mutations. Although recurrence of mutation of the same genomic position across cancer patients has been used as a main feature of driver mutations this is clearly not sufficient. Additionally, in contrast to coding regions of the genome in which the phenotypic changes can be predicted by the genetic code that matches each trinucleotide to its associated amino acid, in non-coding regions this is not usually the case; the reason is that we do not fully understand the functions of non-coding regions. Additionally, non-coding mutations might confer selective advantages that are subtler than those in coding sequences which can induce aberrant protein production and gene silencing, which are easier to identify and explore with downstream experiments and further analyses.

Declines in costs associated with whole genome sequencing have increased the number of cancer genomes available and the analysis of recurrent mutations has been potentiated to identify potential driver mutations. Here it is demonstrated that using mutation recurrence alone it is not sufficient to characterise non-coding mutations. It is shown that recurrent mutations are more enriched in non-B DNA motifs, which is most likely the result of hypermutation at those loci. A previous observation indicated that an inverted repeat at PLEKHS1 promoter is hypermutable at the spacer sequence with no clear phenotypic effects (Lawrence et al. 2014), (Nik-Zainal et al. 2016). Similarly, a set of inverted repeats with a particular spacer sequence were found to be recurrently mutated across cancer genomes (Nik-Zainal et al. 2016), (Zou et al. 2017).

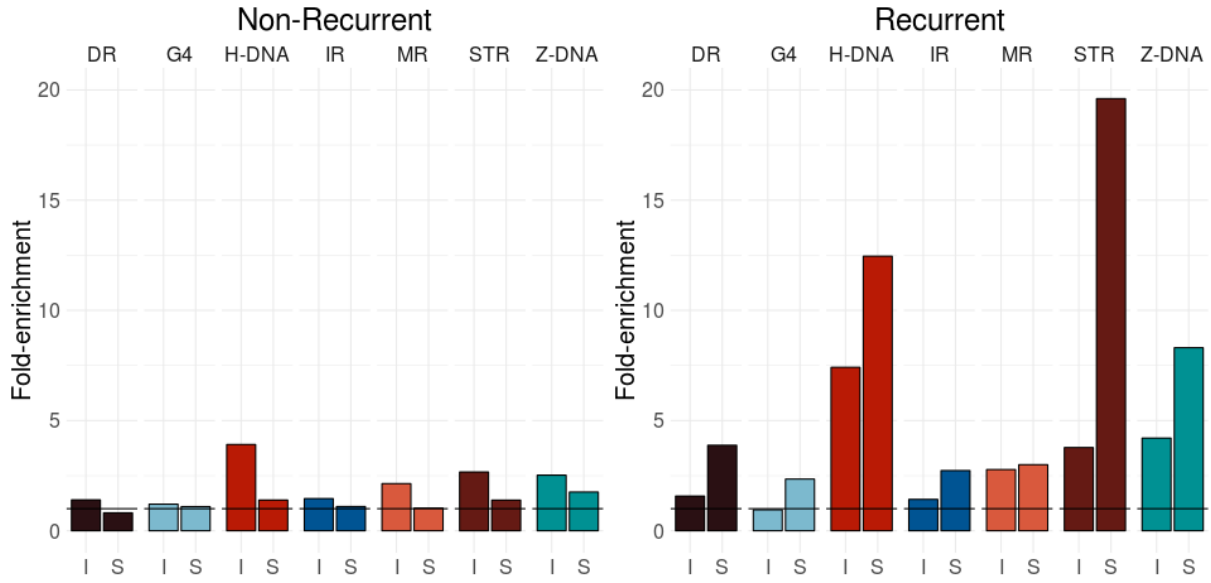
As a result, this analysis builds on these previous findings to suggest that non-B DNA motifs have a higher chance of being recurrently mutated, therefore obfuscating recurrent mutagenesis in identification of driver mutations. The number of mutations at each position in the genome was calculated with a custom python script. Recurrently mutated positions in the genome were observed mutated in multiple independent patients, while non-recurrent mutations were the vast majority (Figure 3.5a) and were not shared

between patients. It was shown that recurrent mutations are more frequent than expected by change in cancer genomes, using a truncated Poisson distribution (Figure 3.5a). In addition, recurrent indels and substitutions were found more often overlapping non-B DNA motifs than non-recurrent indels and substitutions in breast cancer (Figure 3.5b-c). Results for the other cancer types were consistent (Georgakopoulos-Soares et al. 2018).

These results implicate non-B DNA motifs in driver mutation identification. As a result, future genome-wide statistical models of mutability should take into consideration the role of non-B DNA sequences in promotering mutagenesis.



a.



b. c.

Figure 3.5: Non-B DNA motifs contribute to locally elevated mutation rates resulting in recurrent mutations in the human genome across cancer types.

a. Distribution of the number of recurrent events for 3,476,890 somatic mutations from 560 breast cancers. The values do not fit a truncated Poisson distribution (Chi2-test, p<1e-16) as there are more recurrent mutations than predicted by the null model. b. Enrichment of non-recurrent mutations overlapping non-B-DNA motifs for indels (I) and substitutions (S). c. Enrichment of recurrent mutations overlapping non-B DNA motifs for indels (I) and substitutions (S). Mann-

Whitney U test for substitutions: p-value <0.001 for all non-B DNA motifs. Mann-Whitney U test for indels: p-value <0.001 for short tandem repeats (STRs), H-DNA, Z-DNA, mirror repeats (MR) and p-value <0.05 for direct repeats (DR) and G-quadruplex (G4).

3.5. Discussion.

Non-B DNA motifs can be predictors of the distribution of mutations in the human genome across different cancer types. Models combining epigenetic modifications, replication timing and non-B DNA motifs perform better than models with epigenetic modifications and replication timing or non-B DNA motifs separately. Because of the availability of non-B DNA motifs by the primary DNA sequence, their application to cancer types independent of cell of origin and the associated increase in performance, future studies that predict mutational patterns in the genome should implement non-B DNA motifs as features in the models of genome-wide mutability.

The fraction of the genome covered by non-B DNA motifs is less than 10% (Figure 2.2a). Thus, even though these elements are highly enriched for mutations (Figure 2.6) and in particular recurrent mutations (Figure 3.5c), this still only represents a minority of all mutations and is the reason why negative correlations were also observed for certain non-B DNA motifs (Figure 3.1c-d). To demonstrate that this peculiarity is a consequence of scale, an analysis was performed examining non-B DNA motif mutability at much smaller windows of 2kb instead of 500kB (Figure 2.6), centering on substitutions / indels and observing the distribution of each non-B DNA motif, from which a very clear correlation to mutability for non-B DNA motifs was observed.

Furthermore, not all occurrences of a non-B DNA motif are equally mutable. The mutational profile of each category of non-B DNA motifs displays variance which relates to sequence characteristics such as nucleotide composition, spacer and arm lengths for direct repeats, mirror repeats and inverted repeats and average loop size for G-quadruplexes. Within each of those structures the mutational density is unevenly

distributed with regions that are more exposed harbouring an excess of mutations (Figure 3.3-3.4, Figure 3.6).

Finally, recurrent mutagenesis is enriched at non-B DNA motifs, implicating them in deciphering passenger mutations in highly mutagenic regions of the genome from the subset of recurrent mutations that confer selective phenotypic advantages. Importantly, the vast majority of disease-causing variants are found in non-coding regions of the human genome (Maurano et al. 2012) but their functional effects remain difficult to elucidate in most cases. Presence of non-B DNA motifs at non-coding, putative driver variants should be included in future models that search for driver mutations and a negative weight should be added for mutations overlapping non-B DNA motifs (Figure 3.6). The best described example of a genomic site that is hypermutable is the inverted repeat at PLEKHS1 promoter, for which as of now there is no conclusive evidence for any selective advantages, but is nevertheless recurrently mutated across multiple cancer types. However, it should be noted that there could be cases in which a mutated non-B DNA motif has an effect in cancer progression, as experimental evidence suggests functional, regulatory roles in the genome. To date, such examples have not been presented from analyses of recent WGS cancer consortia.

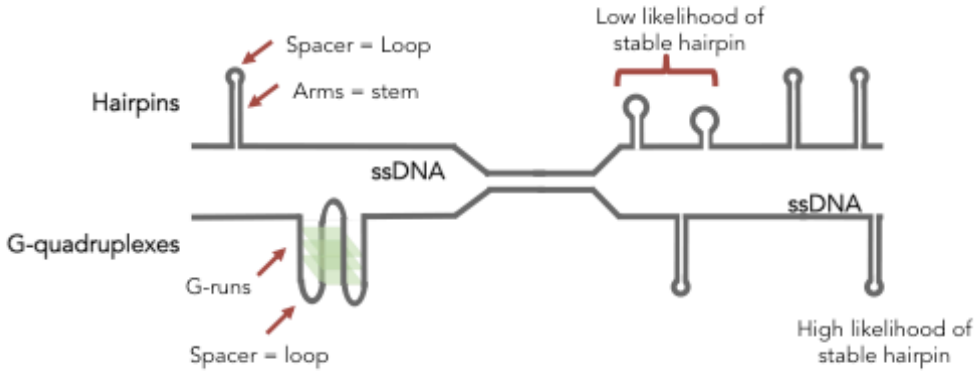


Figure 3.6: Schematic representation of hairpin and G-quadruplex formation along the DNA molecule.

Sequence characteristics such as arm length and spacer size are implicated in the likelihood of secondary structure formation and its stability. Differences in mutability in the subcomponents stem from the exposure of single stranded DNA at loops and spacer sequences.

CHAPTER FOUR

4. Homologies and non-B DNA motifs at indel sites in cancer genomes.

In this chapter, a characterization of factors influencing indel mutagenesis is performed across multiple cancer types. Non-B DNA motifs are enriched for indels and there are patterns of enrichment specific to the type of indel (insertion, repeat-mediated deletion, michromology-mediated deletion) and non-B DNA motif categories. Sequence homologies contribute to differences in indel mutagenesis. More specifically, inserted sequences display higher sequence similarity to the proximal insertion site in comparison to deleted sequences. Furthermore, insertions and deletions display distinct patterns of mutagenesis, suggesting that they result from different DNA damage and repair mechanisms.

4.1. Introduction: Characterisation of indels in cancer genomes.

Indels represent the second most common type of mutation, following substitutions, with important pathogenic implications. However, their role in cancer has remained substantially less studied than that of substitutions (Imielinski et al. 2017), partially due to higher false positive rates. Patterns of substitutions have been thoroughly analysed and their genomic and epigenomic mutational landscape has been described in detail (Nik-Zainal et al. 2012a), (Alexandrov et al. 2013), (Polak et al. 2015), but similar systematic and comprehensive analyses have not been performed for indels to date.

Deletions can be classified by the mechanism responsible for their formation into repeat-mediated and microhomology-mediated (Nik-Zainal et al. 2012a), therefore providing

insight into the mutational processes that contribute to their formation in a cancer patient genome. Microhomology-mediated deletions are enriched in patients with deficiencies in homologous recombination (HR). HR is an error-free mechanism for repair of double strand breaks (Li and Heyer 2008), (Jasin and Rothstein 2013). In its absence cells use alternative repair pathways including non-homologous end joining (NHEJ), which is an error prone mechanism of double-strand DNA repair. Microhomology-mediated deletions tend to be larger than 3bp and show a pattern of homology in the immediate vicinity of the deletion (Nik-Zainal et al. 2016). On the other hand, repeat-mediated deletions are enriched in Mismatch repair deficient (MMR-deficient) tumours. Repeat-mediated deletions tend to be smaller or equal to 3bp and a repeat pattern is observed in the immediate vicinity of the deletion site (Richman 2015).

Similarly, to substitutions, the distribution of indels along the human genome is not random. For instance, nucleosome positioning and replication timing are associated with differences in indel distribution patterns along the genome (Morganella et al. 2016). In addition, the proximal sequence context at the indel site has also been implicated in their formation (Tanay and Siggia 2008). Indels have been previously associated with the presence of non-B DNA motifs both with bioinformatic analyses and in experimental studies (Wang and Vasquez 2006), (Kurahashi et al. 2009), (Wojcik et al. 2012), (Damas et al. 2012), (Lu et al. 2015), (Kamat et al. 2016), (Bacolla et al. 2016), (Zou et al. 2017), (Georgakopoulos-Soares et al. 2018). Nevertheless, a detailed examination of non-B DNA motifs at insertions, microhomology-mediated deletions and repeat-mediated deletions across cancer types or cancer genomes has not been performed to date.

Here, indels from 2,575 tumours derived from 21 organs (Table 4.1) were characterised. It was found that non-B DNA motifs overlap a large portion of indels, there are non-B DNA motif-specific mutational patterns and their sequence characteristics influence the likelihood of indel mutagenesis. Finally, sequence homologies were discovered and characterised at indel sites and the association between indel categories and regulatory elements was explored.

Table 4.1: Number of patients, insertions and deletions by tumour organ.

Tumour organ	Patients	Deletions	Insertions
Bladder	23	11,101	5,571
Biliary	34	112,952	35.024
Pancreas	313	93,936	91,392
Head / Neck	56	23,756	14,602
Liver	314	150,392	78,977
Ovary	110	59,917	27,903
Prostate	199	36,017	22,512
Colorectal	52	208,761	132,204
Myeloid	38	1,177	609
Stomach	68	253,355	62,045
Cervix	20	3,854	3,434
Uterus	44	119,848	78,578
CNS	287	29,362	19,497
Lymphoid	197	61,209	43,592
Skin	107	79,358	27,657
Kidney	186	104,359	29,518
Breast	211	70,333	23,088
Esophagus	97	89,741	63,642
Thyroid	48	3,101	1,045
Bone	89	14,256	4,527
Lung	82	89,842	34,210

4.2. Indel variant calling and distribution patterns at cancer genomes.

Indels have traditionally resulted in a higher false discovery rate than substitutions. In the described analysis, somatic indel calls were performed using three pipelines from four somatic variant callers, therefore decreasing the number of false positives from the mutational calling process. These were the Wellcome Sanger Institute pipeline, the DKFZ/ EMBL pipeline and the Broad Institute pipeline as described in (Campbell et al. 2017), with combined false discovery rate of somatic variants at 2.5%. Indel calling was performed by those algorithms and only indels called by at least two of the callers were analysed (Campbell et al. 2017), therefore generating a conservative dataset (Table 4.1). Additionally, the false positive rate was lower than in most published cohorts to date.

Here, indels were defined as insertions or deletions ranging in size between one and hundred nucleotides. Across the 2,575 WGS samples, the median number of indels per patient sample was 386, with deletions (median of 222) being more prevalent than insertions (median of 124) in the majority of cancer patients (Figure 4.1a). The ratio of deletions to insertions varied substantially by tumour type, between 0.8 to 4.39 (Figure 4.1b). The median size of indels also varied (Figure 4.1c-d), with deletion size varying notably more in each tumour organ between patients than insertion size (Figure 4.1c-d). This is likely the result of distinct mutational processes generating deletions and insertions; the majority of insertions are likely the result of replication strand slippage events, whereas HR and MMR deficiencies influence the median size of deletions per patient towards larger and shorter deletion sizes respectively.

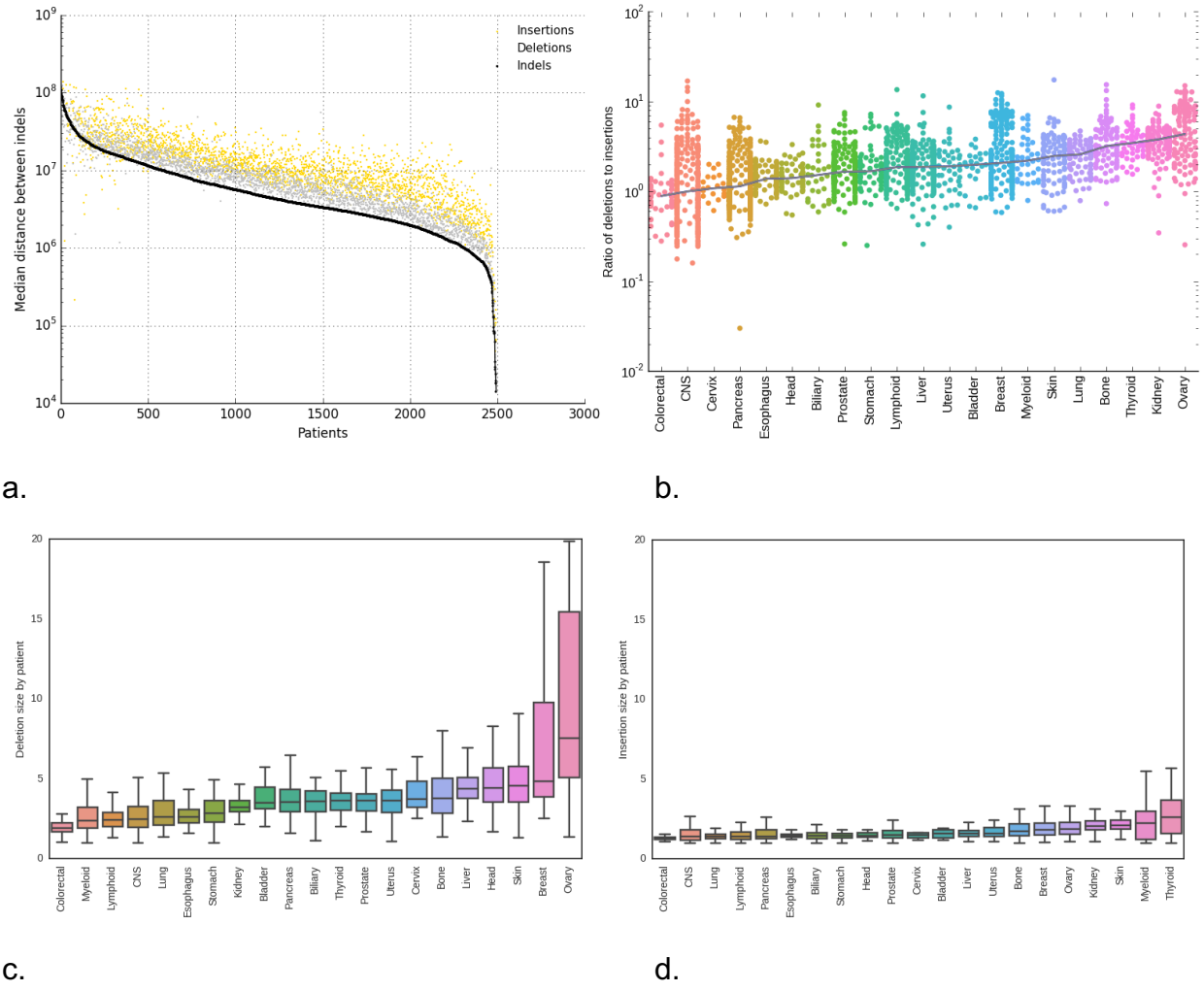


Figure 4.1: Features that influence the frequency and type of indels across cancers.

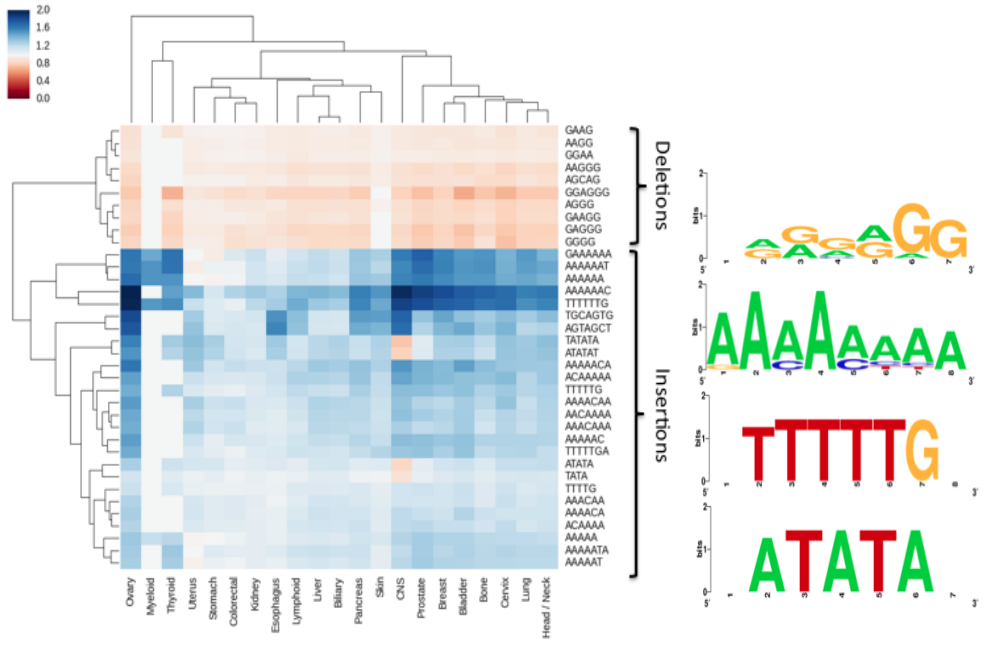
a. Median distance between consecutive indels by patient across tumour types. Separate analysis of insertion and deletion consecutive distances in yellow and grey. b. The ratio of deletions to insertions for each tumour type is shown. c. Distribution of deletion size across patients by tumour type. d. Distribution of insertion size across patients by tumour type.

4.3. Sequence determinants of indel formation.

It was hypothesized that differences in the sequence information in the vicinity of insertion and deletion sites could reflect preferences for insertion and deletion mutagenesis events respectively. Towards that aim, a custom script was developed to investigate if kmer

motifs were found more frequently at insertion or deletion sites, than would be expected by chance. At a local window (+/-150bp) around each insertion or deletion, the density of every monomeric to heptameric motif (21,844 motifs) was calculated using the reference human genome, separately for each organ. If a motif was not biased towards insertions or deletions, then its frequency should be similar at insertion and deletion sites at that organ (after correcting for differences in insertion and deletion frequencies in each organ). However, differentially enriched motifs that favoured insertions or deletions should be present at higher density at insertion or deletion sites respectively.

Indeed, most motifs were not differentially enriched between insertions and deletions. Nevertheless, a small subset of kmer motifs showed preference for insertions or deletions in each organ. As expected, the number of differentially enriched motifs between insertion and deletion sites decreased rapidly, if it was required for the same motif to be differentially enriched in multiple organs. Therefore, only motifs reported as differentially enriched in at least 18 of 21 tissues at insertion or deletion sites were selected, since these were likely to reflect stronger differences between insertions and deletions. Two main clusters were identified, one for insertions (blue) and one for deletions (red) (Figure 4.2a), representing motifs that favoured the formation of insertions or deletions respectively across a multitude of cancer types, therefore implicating the sequence context in the likelihood of insertion or deletion formation. Finally, the similarity between the enriched motifs was calculated and they were clustered using k-means clustering and CLUSTALW alignment. In particular, the deletion cluster was dominated by a single group of AG-rich motifs, whereas three motif groups were identified for insertions (Figure 4.2b). Finally, the three deletion motif groups resembled repeat sequences (Figure 4.2b) and were least differentially enriched at tumours with MMR-deficient samples, such as in colorectal tumours (Figure 4.2a).



a. b.

Figure 4.2: Sequence determinants of insertion and deletion formation.

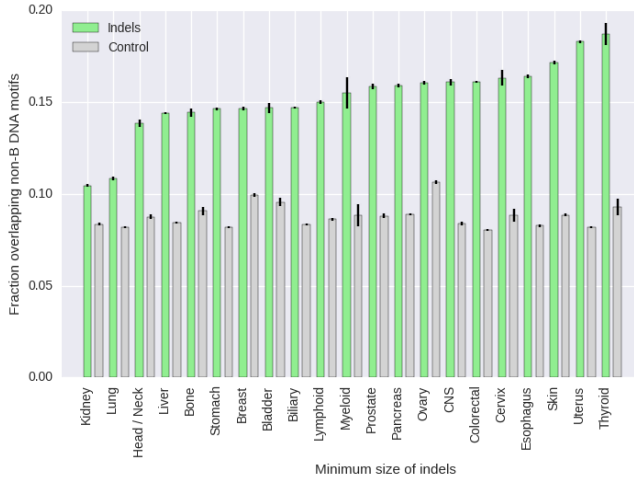
a. Comparative motif preference in the background reference sequence where the insertions and deletions occurred. Enrichment is calculated as the ratio of insertions to deletions. Differentially enriched deletion motifs tend to be AG-rich. Only motifs that were differentially enriched across all tumour types but three or less were used. b. Position Weight Matrices (PWMs) representing motifs that were preferentially enriched at insertions or deletions.

4.4. Analysis of non-B DNA motifs at indel sites.

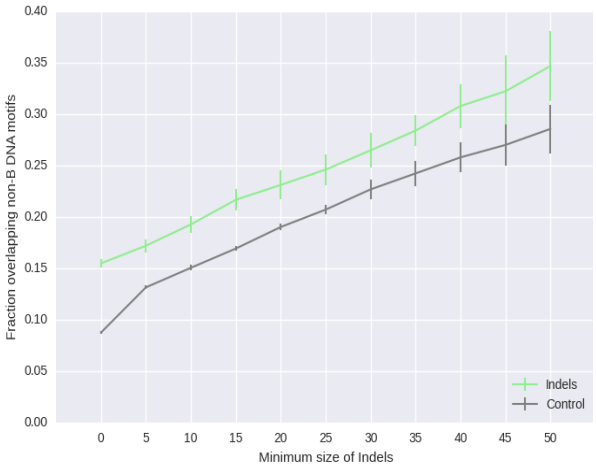
The mutational patterns of indels were investigated relative to non-B DNA motifs. Firstly, the enrichment levels of non-B DNA motifs at indels were explored across the 21 organs. It was found that on aggregate, indels were enriched at non-B DNA motifs across all organs, but the level of enrichment was nearly two-fold different between disparate organs (e.g. Kidney-1.25-fold and Uterus-2.24-fold) (Figure 4.3a). In addition, the enrichment at non-B DNA motifs was increased across all indel sizes (Figure 4.3b). These results reflected the higher propensity of indels to form at non-B DNA motifs. As a next step, individual differences at each non-B DNA motif were investigated.

When subdivided by the different categories of non-B DNA motifs, distinct patterns between indel size and enrichment for each non-B DNA motif were identified. It was found that IRs were enriched for indels of longer lengths (Figure 4.3c). In contrast, short indels <5bp were particularly enriched at short tandem repeats, Z-DNA, H-DNA and mirror repeats, <10bp indels were enriched at G-quadruplexes and 5-25bp indels were particularly enriched at direct repeats (Figure 4.3c). These results reflect non-B DNA motif differences related to indel mutagenesis.

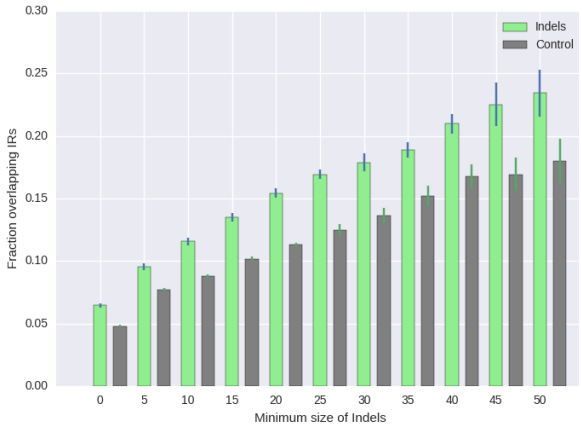
Both insertions and deletions were enriched across non-B DNA motifs (Figure 4.3d), reinforcing earlier observations (Figure 4.3a), although the level of enrichment varied between non-B DNA motifs (Figure 4.3d). Based on the mutational processes that have been reported to underpin deletion formation, deletions were subdivided into repeat-mediated and microhomology-mediated, both of which are defined by the flanking sequences at deletion junctions. Repeat-mediated deletions were enriched in all types of non-B DNA motifs, presumably indicating the relative ease with which small insertion-deletion loops are formed (Figure 4.3e). By contrast, microhomology-mediated deletions were selectively enriched at H-DNA, mirror repeats, inverted repeats and direct repeats and depleted at short tandem repeats and G-quadruplexes. Enrichment of microhomology-mediated deletions at non-B DNA motifs likely reflects double strand break formation during resolution of non-B DNA secondary structures. To gain further insight, an analysis of a subset of non-B DNA motifs by their physical properties followed.



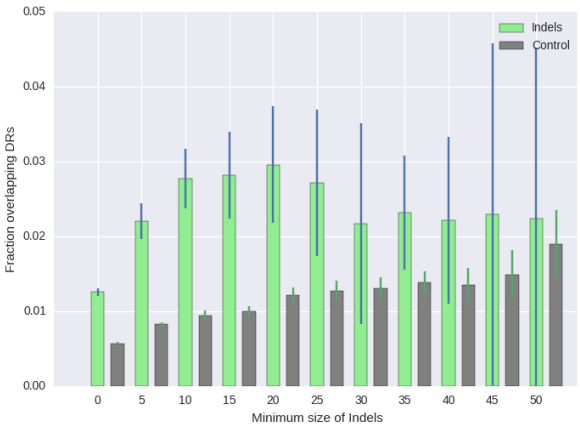
a.



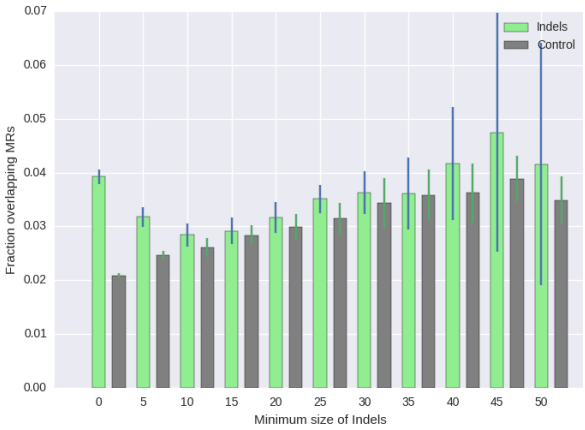
b.



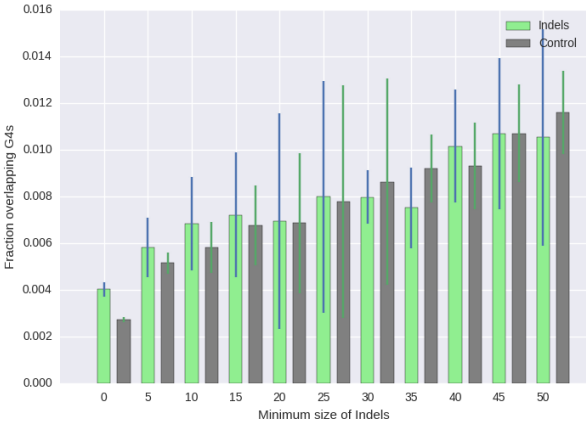
Inverted Repeats



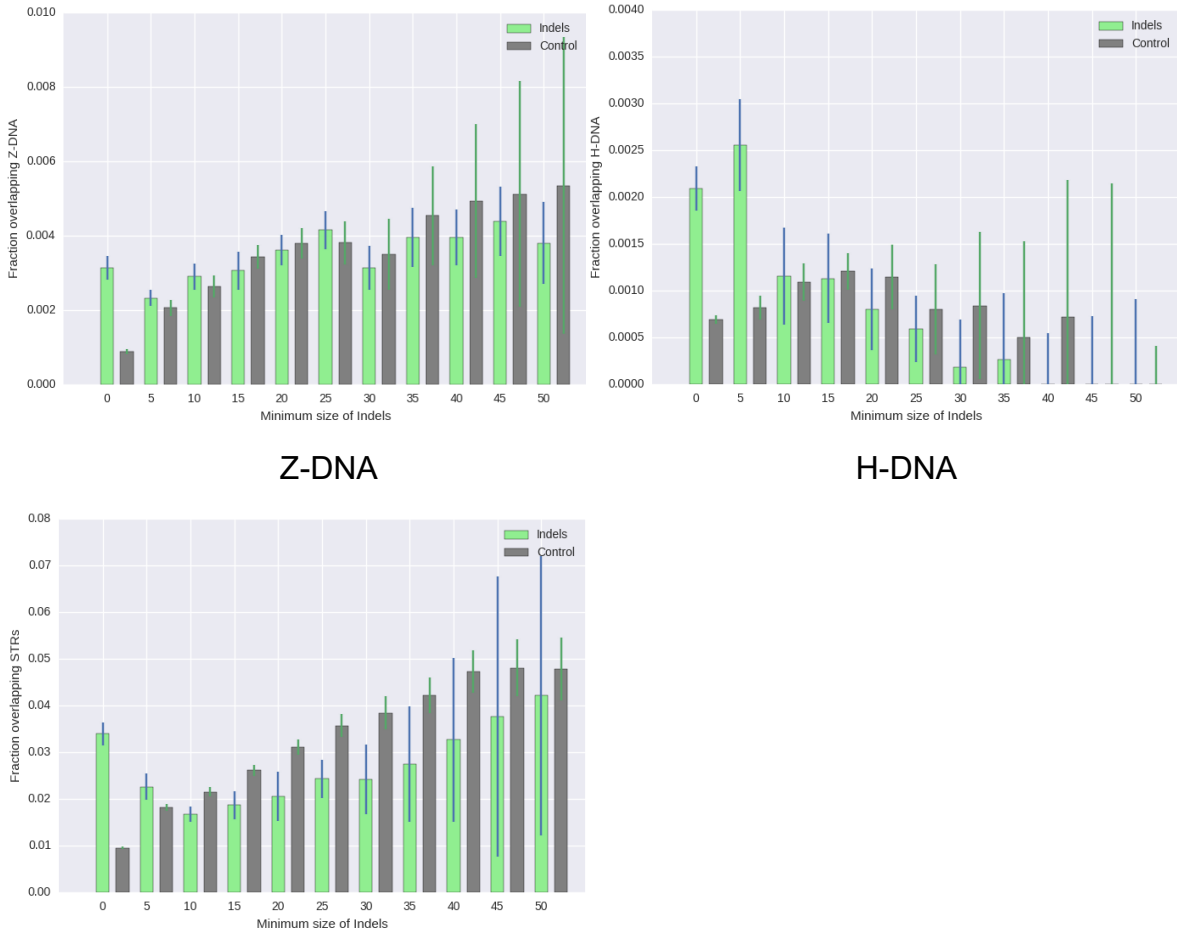
Direct Repeats



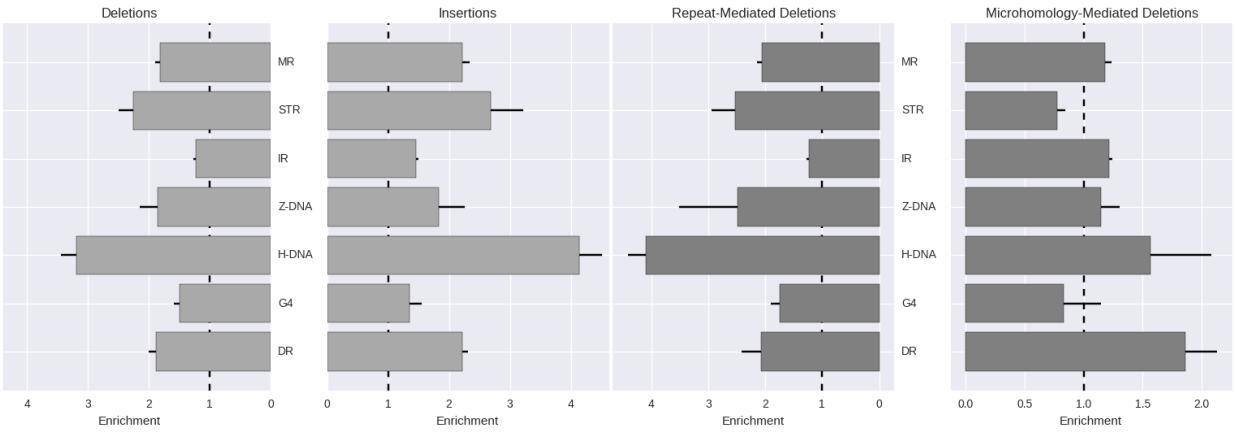
Mirror Repeats



G-quadruplexes



c.



d. e.

Figure 4.3: Enrichment of indels at non-B DNA motifs.

a. Proportion of indels overlapping any non-B DNA motif by tumour type. b. Enrichment across all non-B DNA motifs versus controls for indels by minimum size of indels (cut-off), data of median across tumour types shown with standard error. c. Non-B DNA motif-specific enrichment. d. Enrichment of non-B DNA motifs at insertions and deletions. Error bars represent standard error across tumour types. e. Enrichment of non-B DNA motifs at indel signatures. Error bars represent standard error across tumour types. For panels a-c controls were simulated for each indel controlling for genomic location by placing the indel randomly 500bp away from the original indel site. In panels (d-e) MR is mirror repeats, STR is short tandem repeats, IR is inverted repeats, G4 is G-quadruplex and DR is direct repeat.

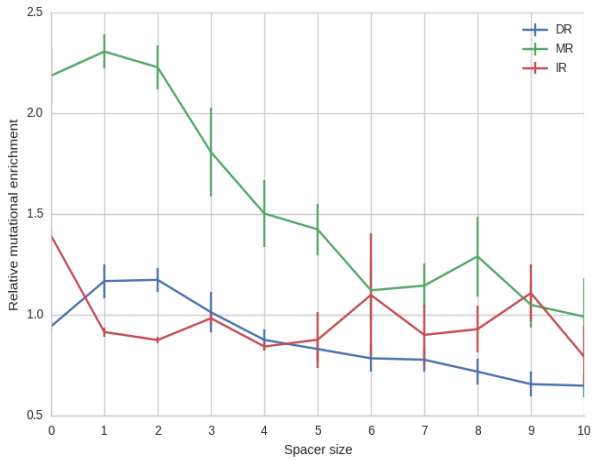
4.5. Sequence characteristics of non-B DNA motifs influence indel mutability.

Previous studies have indicated that mutational density at non-B DNA motifs is influenced by their sequence characteristics (Nik-Zainal et al. 2016), (Zou et al. 2017), (Georgakopoulos-Soares et al. 2018). Physical features of non-B DNA motifs influence the likelihood of their formation and their stability (Varani 1995), (Nag and Kurst 1997), (Woodside et al. 2006), (Tippana et al. 2014), (Piazza et al. 2015). In this section it was explored whether physical properties of non-B DNA motifs, such as the spacer length, contribute to differences in indel mutability. In addition, it was hypothesized that mutability at non-B DNA motifs will be associated with specific mutational processes and separate analysis was performed focusing on insertions, microhomology-mediated deletions and repeat-mediated deletions. Indeed, striking differences were observed at distinct indel categories.

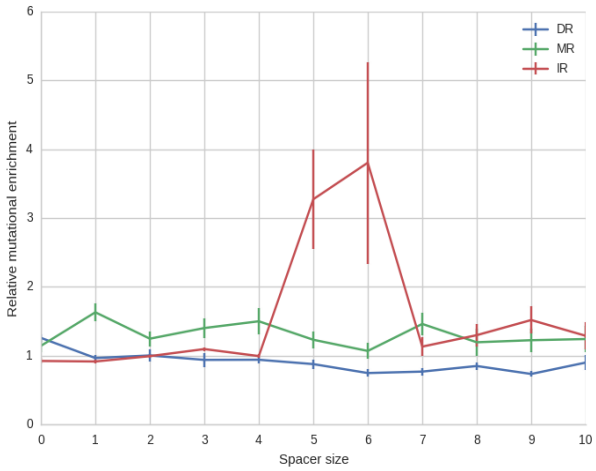
The mutational densities of inverted repeats, mirror repeats and direct repeats were investigated as a function of spacer size, for spacer sizes of 0bp to 10bp. The subset of direct repeats with spacer size smaller than 1bps are more mutable than the rest, for repeat-mediated deletions (Figure 4.4d). This favours a model of slipped structure formation during replication (Garcia-Diaz and Kunkel 2006), (Gadgil et al. 2017), which is more likely to occur at direct repeats with smaller spacer lengths, because they are more likely to fold in slipped structures (Pearson et al. 1998).

Moreover, it was found that inverted repeats with a spacer size between five and six base pairs are almost four-fold more mutable for deletions as other inverted repeats, but not for insertions (Figure 4.4a-b). The increased mutational enrichment was pronounced in microhomology-mediated deletions but not in repeat-mediated deletions (Figure 4.4c-d). Additionally, in contrast to the majority of inverted repeats which have similar spacer to arm indel mutability, in inverted repeats with spacers of length 5-6bps, arms are much more likely to harbour microhomology-mediated deletions, which is not the case for repeat-mediated deletions (Figure 4.4e-f). Previous experimental studies have suggested that inverted repeats with spacer size of 5-6bps are more likely to form hairpins (Varani 1995) and this may in turn result in replication stalling (Voineagu et al. 2008), (Lu et al. 2015). These results support a model of double strand break formation at hairpin arms, preferably at inverted repeats of five or six nucleotides spacer, which fail to be resolved and repaired effectively, resulting in the loss of inverted repeat arms and the formation of microhomology-mediated deletions.

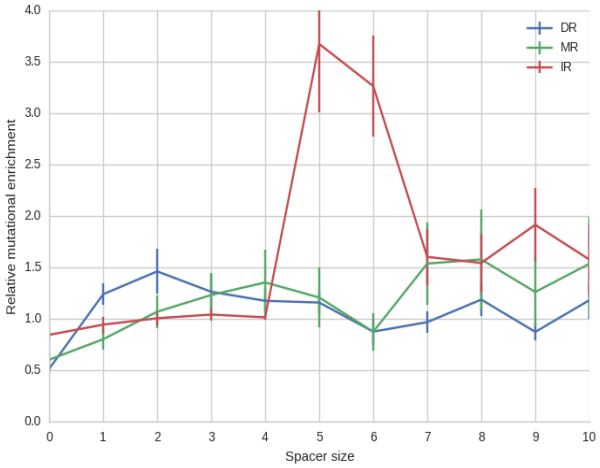
Finally, for mirror repeats, a pronounced enrichment for insertions at mirror repeats with small spacer sizes was observed, with a rapid decrease with increased spacer length, which was not observed in deletions (Figure 4.4a-b). The mutational enrichment for particular spacer sizes for specific indel types suggests that specific sequence characteristics influence the mutability of non-B DNA motifs.



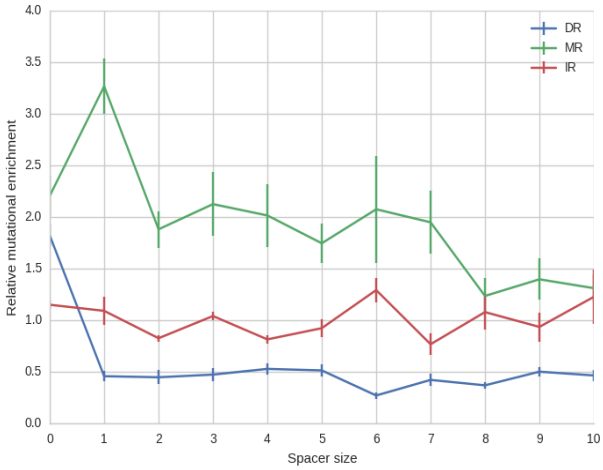
a. Insertions



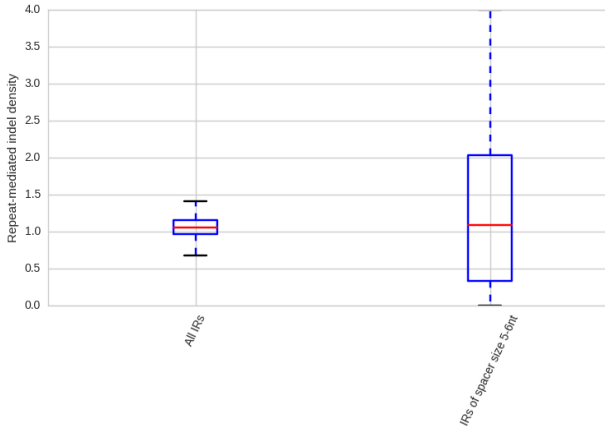
b. Deletions



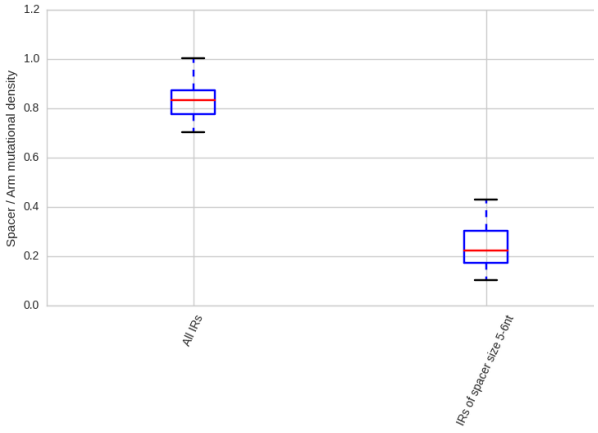
c. Microhomology-mediated deletions



d. Repeat-mediated deletions



e. Repeat-mediated deletions



f. Microhomology-mediated deletions

Figure 4.4: Sequence characteristics of non-B DNA motifs that influence the likelihood of insertions and deletions.

Enrichment of a. insertions and b. deletions in direct repeats, mirror repeats and inverted repeats by spacer length. Enrichment of c. microhomology-mediated deletions and d. repeat-mediated deletions at direct repeats, mirror repeats and inverted repeats. For a-d panels, relative enrichment is calculated as the ratio of the mutational density for a particular spacer size over the average mutability of the non-B DNA motif for the particular tumour type and standard error is shown. e. Box-plot comparing spacer-arm mutability for repeat-mediated deletions. f. Box-plot comparing spacer-arm mutability for microhomology-mediated deletions.

4.6. Sequence similarities and homologies at indel sites.

Population studies have shown that the sequence context found at indel sites is implicated in their formation in both flies and primates (Tanay and Siggia 2008). Thus, it was hypothesized that flanking regions at the indel site are implicated in indel formation, and it was investigated if indels tend to show different types of homologies to their flanking regions. Homologies between the indel sequence and the vicinity where it was formed were investigated, focusing on direct, inverted and mirror symmetries.

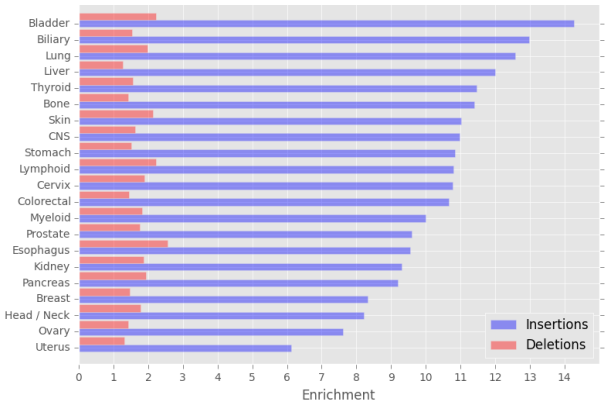
For insertions and deletions >5bp in length, their motif was used to investigate for the presence of the same sequence (direct symmetry analysis) in the indel vicinity (flanking 500bps on either side), using as controls the shuffled background sequence and a sequence 2kB away. A staggering enrichment was found for insertions; in the majority of cases the inserted sequence was also present in the genomic window, typically in the immediate vicinity of the insertion site. These mini tandem repeats are likely the result of replication slippage, where the same segment of DNA is replicated twice, resulting in an insertion with the same sequence. Replication slippage at insertions has previously been suggested by evolutionary studies (Messer and Arndt 2007). A deleted sequence was also likely to have a second copy in the vicinity but the enrichment was substantially smaller (Figure 4.5a). Following the direct symmetry analysis, it was investigated if the inverted or mirror indel sequence could be identified in the indel vicinity. For both inverted and mirror symmetries there was an enrichment, which varied by tumour type (Figure 4.5b-c), although it was not as strong as the enrichment for insertions (Figure 4.5a).

The median fraction of insertions >5bp, for which the same sequence, its mirror, or inverted sequence was found in the background was 86.9% across cancers, with the vast majority found creating novel “mini” tandem repeats (85.5%, 8.4%, 13,1% for direct, inverted and mirror symmetries). This was clearly not the case for deletions. For deletions >5bp, the median frequency across tumours to destroy a direct, inverted or mirror repeat i.e. to find the same motif of the deletion or its reverse complement or its mirror motif, was 10.91% (7.1%, 3.8%, 4.4% for same, inverted and mirror symmetries). It was concluded that there are homologies between the inserted and deleted sequences and the site of indel formation and predominantly direct homology at insertion sites.

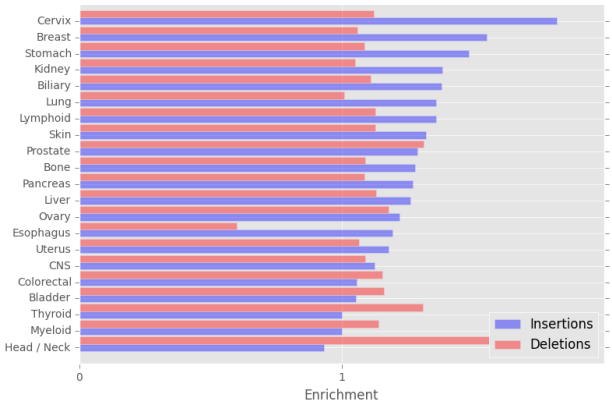
To further investigate sequence homologies at indel sites, the average Hamming distance between insertions or deletions of different minimum size limits and the genomic indel site was explored. The average Hamming distance between a deleted sequence and the background deletion site increased rapidly with increased deletion size, as expected (Figure 4.5d). However, this was not the case for insertions, for which the average Hamming distance remained largely unchanged, as the minimum insertion size increased (Figure 4.5d), suggesting that the inserted sequence was usually found at the insertion site.

Next, the analysis was focused specifically on indels >=10bp. Hamming distance is a similarity measure; it indicates how many positions between two equal sized sequences are different. A Hamming zero of zero indicates the same sequence. Interestingly, the Hamming distance for insertions was much lower than for deletions (Figure 5e-f), and in the majority of cases the inserted sequence was also present in the indel site (85.3%), therefore reinforcing earlier observations (Figure 4.5a). The high degree of similarity between the inserted sequence and the site of insertion implicates replication slippage in the formation of the majority of insertions (Messer and Arndt 2007), as also suggested from earlier analysis (Figure 4.5a). Finally, the Hamming distance between the mirror or inverted sequence of the inserted / deleted motif and the indel site was calculated to investigate other potential homologies being present. The differences between insertions

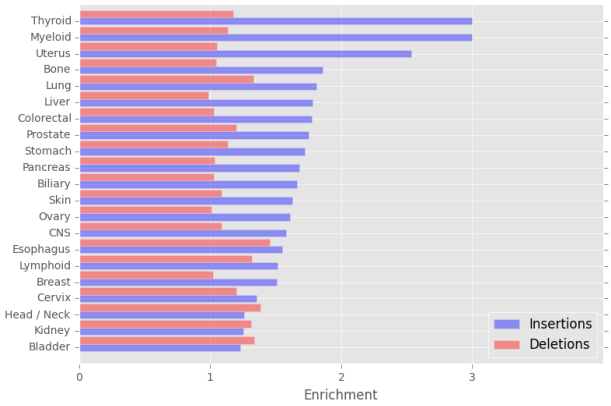
and deletions at inverted and mirror sequences were not comparable to those observed for the motif itself, as also indicated from earlier analysis (Figure 4.5a-c).



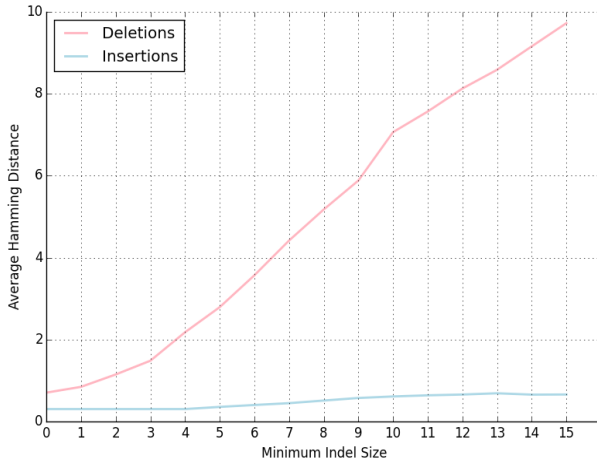
a.



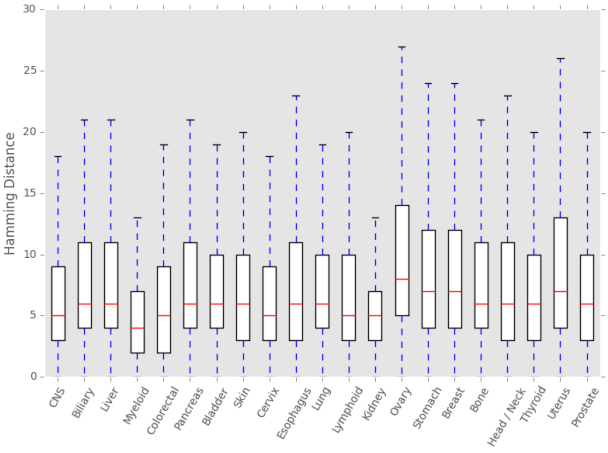
b.



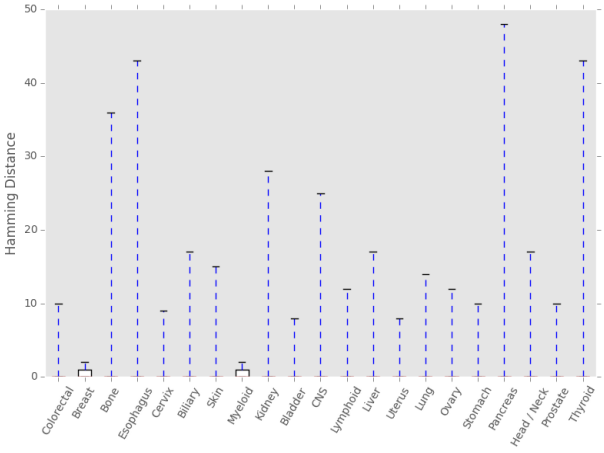
c.



d.



e.



f.

Figure 4.5: Homology patterns at indel sites.

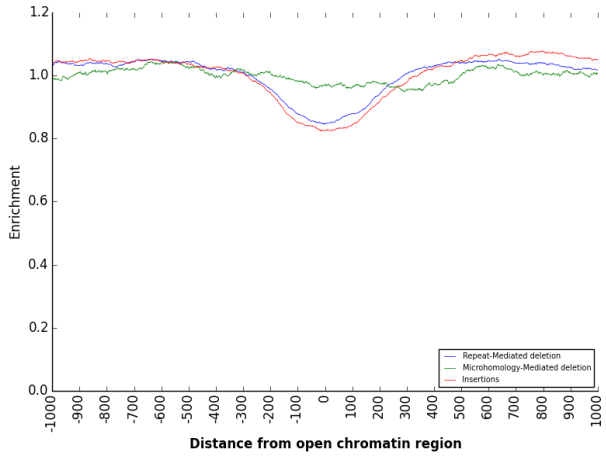
Enrichment of motifs in the surrounding sequence from where the indel occurs. a. Same sequence, b. Mirror sequence, c. Inverted sequence. Indels> 5bp shown. Enrichment is the ratio of the observed versus expected occurrences of inserted / deleted sequence being found in the window around the site of insertion / deletion. Expected occurrences are calculated using two controls. d) Average Hamming distance across tumour types for insertions and deletions, across a range of minimum indel sizes. eHamming distance between e. deletion of >=10bp and background window (left) and f. between insertion of >=10bp and background window (right). [For deletions the deleted sequence was excluded from the search space].

4.7. Associations between indel categories and regulatory elements.

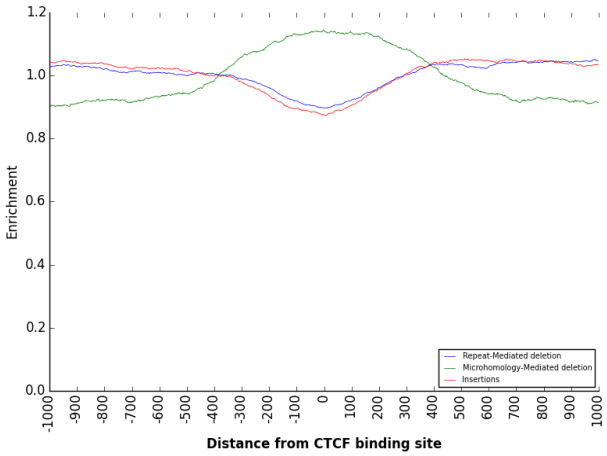
DNA damage and repair has been shown to be influenced by the local chromatin landscape. Interestingly, open chromatin regions have a lower density of substitutions (Polak et al. 2015), which is likely the result of higher repair accessibility. Transcription factor binding sites (TFBSs) within open chromatin regions display an elevated density of substitutions because bound transcription factors interfere with repair machineries at the damaged sites (Sabarinathan et al. 2016). However, these analyses did not focus on the interplay between regulatory sites and indel mutagenesis. Here, the association between the different indel categories and a set of cis-regulatory elements, including open chromatin, CTCF binding sites and TFBSs was investigated.

On aggregate, repeat-mediated deletions and insertions were de-enriched at open chromatin regions, CTCF binding sites and TFBSs (Figure 4.6a-c), which reflects the preference for heterochromatin, late replicating regions (Morganella et al. 2016). However, it was found that microhomology-mediated deletions were enriched at CTCF binding sites and TFBSs, whereas there was no enrichment at open chromatin regions (Figure 4.6a-c). This was a surprising result, which potentially implicates TFBSs and CTCF binding sites in the formation or repair of double strand breaks, similar to what was previously identified for substitutions (Sabarinathan et al. 2016). As a result, the binding

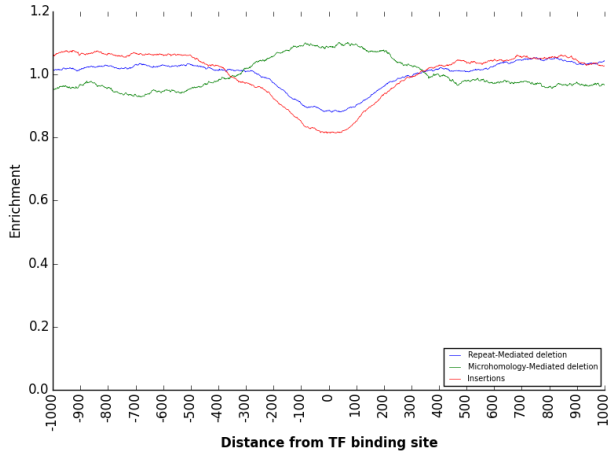
of transcription factors at their cognate sites could influence the likelihood of double strand break formation or its repair as indicated by the enrichment of microhomology-mediated deletions on aggregate across cancer genomes.



a.



b.



c.

Figure 4.6: Associations between indel categories and regulatory elements.

Enrichment of insertions, repeat-mediated deletions and microhomology-mediated deletions relative to: a. open chromatin regions, b. CTCF binding sites, c. transcription factor binding sites. The enrichment score at a position was defined as the frequency of a mutation type at a position relative to the median frequency across all positions within that window.

4.8. Discussion.

A systematic exploration of indels across WGS cancer patients has not been performed to date. Here, the role of non-B DNA motifs, kmer motifs and sequence homologies at indel sites as well as the associations between indel categories and regulatory elements were investigated. The enrichment of non-B DNA motifs at indel sites implicates them in their formation. However, the patterns of enrichment are clearly different for non-B DNA motif and indel categories (Figure 4.3). In addition, tandem repeats have been previously associated with higher rates of false positive mutation calling. As a result, although only a subset of tandem repeats that can be identified with current methods were included in the analysis, results regarding short tandem repeats should be treated with more caution.

Furthermore, it was shown here that the spacer sequence of inverted repeats, mirror repeats and direct repeats is another determinant of their likelihood of indel mutagenesis. For instance, the case study of indel enrichment at inverted repeats for different spacer sizes indicated a relative mutational enrichment at spacer sizes of 5-6 base pairs (~3.5-fold), which was specific to microhomology-mediated deletions and was not found in insertions or repeat-mediated deletions. Further analysis indicated that the microhomology-mediated deletions were focused primarily at the inverted repeat arms and not the spacer sequence (correcting for differences in sizes). As a result, the patterns of mutagenesis suggest an interplay that is specific to particular mutational processes and indel types.

Interestingly, it was shown that the vast majority of insertions generate “mini tandem repeats” and in most cases the inserted sequence is already present in the background insertion vicinity (Figure 4.5). In contrast, this is not the case for deletions. Although further work would be required to describe the underlying mechanism, it is suggested that replication strand slippage causes the formation of these mini tandem repeats.

In the next chapter, the landscape of indel mutagenesis was further explored by investigating strand asymmetries during indel formation.

CHAPTER FIVE

5. Transcriptional and replication strand asymmetries at indels across cancer genomes.

In this chapter, a novel method is devised to systematically investigate transcriptional and replication strand asymmetries for indels using polyN motifs. Firstly, the background distribution patterns of each polyN motif are characterised across genic regions and replication deciles. Next, strand asymmetry biases are identified both at template and non-template strands for transcriptional directionality and between leading and lagging strands for replication directionality. Finally, transcriptional strand asymmetries in cancers are estimated, providing evidence for the contribution of TC-NER and MMR machineries in the observed biases.

5.1. Introduction: A method to measure transcriptional and replication strand asymmetries for indels.

Transcription is a tightly regulated process by which different cell types mediate the production of relevant RNA molecules. Transcription involves the recruitment of the RNA polymerase complex, which in combination with other factors and cofactors, binds the template strand (also known as transcribed / non-coding strand) to induce transcription. During the process of transcription, DNA is found in single stranded form, while nascent RNA is produced through multiple rounds of transcription.

If DNA damage occurs prior to or even during transcription, it can stall the progression of the transcriptional machinery along the DNA. Indeed, previous studies have described the link between transcription and genomic instability (Kim and Jinks-Robertson 2012), (Wang and Vasquez 2017), an example being transcription-replication collisions (Sankar

et al. 2016). In particular, multiple DNA repair enzymes are recruited at the damaged DNA site, most notably transcription-coupled nucleotide excision repair (TC-NER), to ensure the fidelity of the genomic information is maintained and transcription is resumed. TC-NER preferentially corrects DNA damage at the template strand and therefore an excess of mutations is found accumulating in the non-template strand.

In contrast to indel mutations, substitutions are commonly oriented by the trinucleotide context at the immediate vicinity of the substitution. Recent analysis of mutational signatures has indicated that certain substitution signatures show transcriptional strand bias (Morganella et al. 2016), (Haradhvala et al. 2016), (Andrianova et al. 2017), (Tomkova et al. 2018), reflecting biases in DNA damage and repair. For instance, TC-NER-associated transcriptional strand bias has been demonstrated in substitution signatures 4 and 7, representing tobacco and ultraviolet light exposure at lung and skin cancers respectively (Alexandrov et al. 2013). In addition, TC-NER activity is correlated to gene expression levels (Hanawalt and Spivak 2008), with highly expressed genes displaying stronger transcription-induced mutational strand bias. Therefore, transcribed regions show an asymmetric pattern of somatic mutagenesis between template and non-template strands.

For indel mutagenesis, previous research has suggested an anti-correlation between indel rate and gene expression levels (Lim et al. 2017). However, strand asymmetry in indels has not been investigated until now due to technical challenges involving the orientation of indels relative to the transcriptional direction. As a result, putative transcriptional strand asymmetries in indels remain unexplored. Here, a readout to measure transcriptional strand asymmetry of indels is proposed. The analysis is focused at polyN motifs (N denoting G / C / T / A) which are known to be highly enriched for indels.

First, polyN motifs were mapped in the human genome for polyN tract lengths between 1bp and 10 bps. This provided a genome-wide map of non-overlapping polyN motifs, i.e. a “GG” motif was not counted in a “GGGG” motif. This served as the basis for the downstream analysis. Next, polyN motifs were oriented relative to the direction of

transcription, separating them into motifs on the template or non-template strands. Using this method, measurable deviations were identified in the background distribution of polyN motifs across the gene length, which to the best of my knowledge had not been characterised before. Finally, the observed and expected indel mutagenesis rate at those sites were calculated, correcting for the background template / non-template frequencies of polyN motifs. Surprisingly, strong transcriptional-strand asymmetries were identified that were specific to particular tumour types and to particular polyN motifs.

If it is assumed that there were no strand preferences during DNA damage or repair, then there should not be any difference in indel mutagenesis at polyN sequences for template and non-template strands. However, similar to substitutions, indels overlapping polyN tracts show transcriptional strand preferences which are specific to particular tumour types and mutational processes. The observed levels of asymmetry are stronger than those previously reported for substitutions (Morganella et al. 2016) and specific mutational processes are directly implicated in the magnitude of the effect.

During replication, the leading strand is synthesized continuously, while the lagging strand is synthesized in pieces, known as Okazaki fragments. Indels occur at a higher rate in late replicating regions (Morganella et al. 2016) and mutational patterns in cancer genomes are strongly influenced by replication timing (Polak et al. 2015) and display replication strand asymmetries (Morganella et al. 2016), (Andrianova et al. 2017), (Tomkova et al. 2018). In particular, certain substitution signatures display preference for leading versus lagging strand, therefore reflecting biases in DNA damage and repair.

As a result, the question arises if replication strand asymmetries could be observed for indels at polyN motifs. Repli-seq is a commonly used method to identify the locations throughout the genome at which nascent DNA is synthesized, through the incorporation of bromouridine triphosphate instead of thymidine (Hansen et al. 2010). By implementing this method, the genome can be divided into replication domains known as replication time deciles and the direction of the replication fork migration can be imputed, resulting in the separation of leading and lagging domains (Morganella et al. 2016).

PolyN motifs were mapped across the replication time deciles to investigate their distribution patterns relative to replication timing. Finally, a detailed analysis of replication strand asymmetry at polyN motifs was performed, to investigate differences between the leading and the lagging strands. In contrast to the transcriptional strand asymmetry patterns, only weak strand asymmetries were observed for replication timing at indels overlapping polyN motifs.

5.2. Distribution of polyN motifs in genic regions.

A custom script was generated in python to map the distribution of non-overlapping, mononucleotide repeat tracts here termed “polyN motifs” of lengths 1-10bp across the human genome. The background distribution of polyN motifs was characterized at genic regions, considering the gene orientation (Figure 5.1a).

First, the frequency of polyN motifs at the template and non-template strands of genes was explored. PolyA motifs were found more frequently at the template strand than on the non-template strand and the magnitude of the observed difference was more exacerbated at longer polyA motifs (Figure 5.1b). In contrast, the densities of polyG motifs on the template and non-template strands were similar for polyG tracts of 1-6bp, and biases were only observed for long polyG tracts (Figure 5.1b).

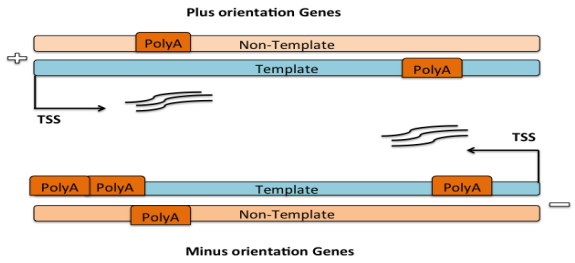
Second, the distribution of polyN motifs was investigated across the gene length. In order to correct for differences in gene length, each gene was divided into ten equal sized bins. Two additional bins were added upstream of the transcription start site (TSS) and two added downstream of the transcription end site (TES), each 10kB in length, resulting in a total of 14 bins. The distribution of polyN motifs across the gene length was heterogeneous. A strong enrichment of polyG motifs was observed at the 5’ and 3’ ends of genes, and directly upstream and downstream of the TSS and TES (Figure 5.1c). In

contrast, polyA motifs were found to be enriched within the body of the gene instead, with a lower enrichment at the last bins of transcribed regions, and depletion upstream of the TSS and downstream of the TES of genes (Figure 5.1c). These results are in accordance with the known GC-skew around the TSS and TES as well as the GC-rich regions at the start and end of genes (Ginno et al. 2013).

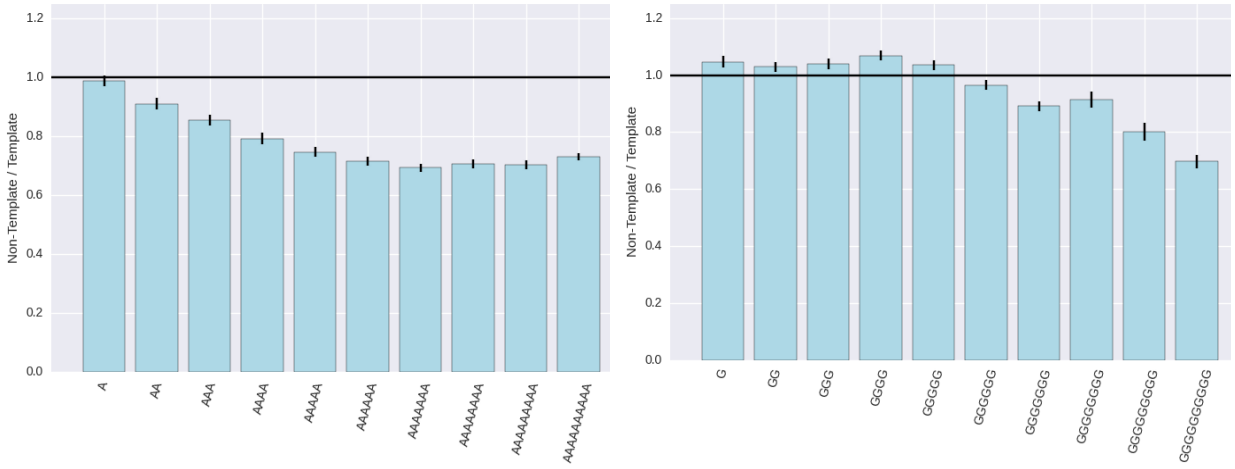
Third, for each polyN motif, the frequencies on the non-template and template strands were explored at each bin across the gene length, to investigate preferences. It was found that polyA motifs displayed a strong bias towards the template strand, which was dependent on the length of the repeat tract and the distance from the TSS (Figure 5.1d). Strong background asymmetry levels were not observed for polyG motifs across the gene length, with the exception of long polyGs (Figure 5.1d), also reinforcing earlier observations (Figure 5.1b).

Fourth, the observed strand asymmetries for polyA and polyG motifs were also investigated relative to the distance from the TSS and the TES, at a nucleotide resolution window. The non-template to template strand difference in polyA motifs was not found upstream of the TSS or downstream of the TES, indicating that the bias is specific to genic regions and is associated with transcription (Figure 5.1f). For polyG motifs, the non-template to template strand asymmetry in frequency was mostly observed in the immediate vicinity of the TSS and TES, it was less exaggerated and the signal disappeared within the next ~2kB, while the peak around the TSS and TES was amplified with increased polyG tract length (Figure 5.1f).

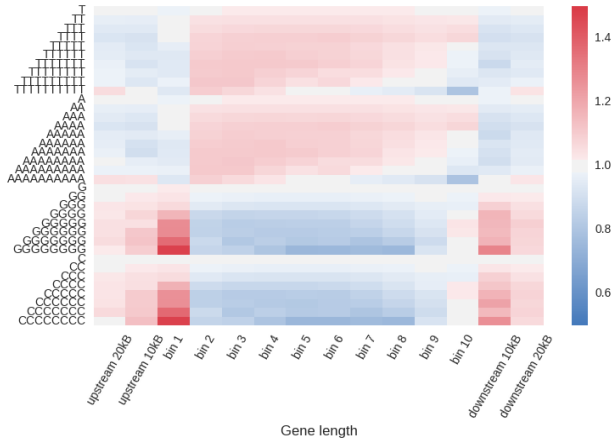
The results described above were very surprising, indicating a clear and (to the best of my knowledge until now) undescribed relationship between the strand preference of polyA motifs in genic regions. In the next section, I investigated whether indel mutagenesis across multiple cancer types displayed a preference for the template or non-template strands, when overlapping polyN motifs.


Plus orientation Genes
PolyA
Non-Template
Template
PolyA
TSS
TSS
PolyA
PolyA
Template
PolyA
PolyA
Non-Template
Minus orientation Genes

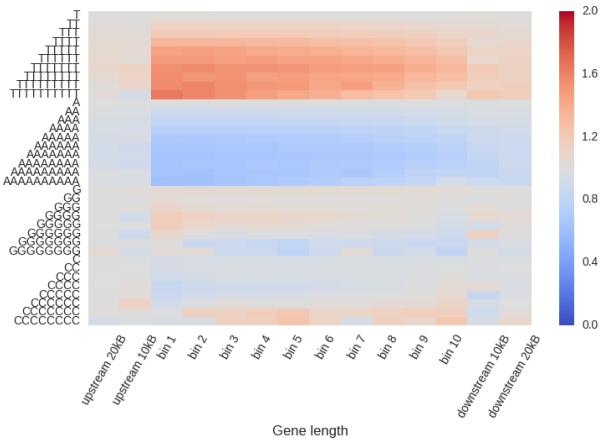
a.


Non-Template / Template
A
AA
AAA
AAAA
AAAAA
AAAAAA
AAAAAAA
AAAAAAAA
AAAAAAAAA
AAAAAAAAAA
Non-Template / Template
G
GG
GGG
GGGG
GGGGG
GGGGGG
GGGGGGG
GGGGGGGG
GGGGGGGGG
GGGGGGGGGG

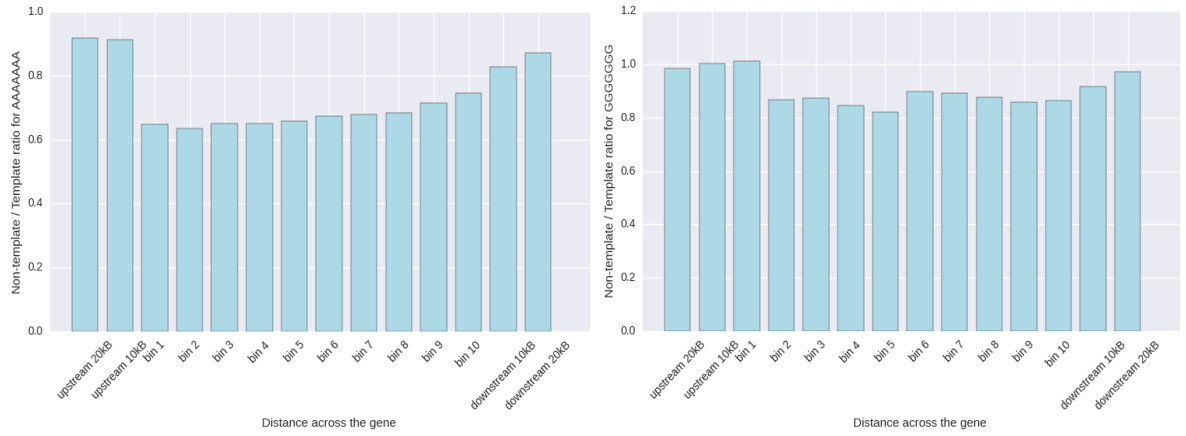
b.


upstream 20kB
upstream 10kB
bin 1
bin 2
bin 3
bin 4
bin 5
bin 6
bin 7
bin 8
bin 9
bin 10
downstream 10kB
downstream 20kB
Gene length

c.


upstream 20kB
upstream 10kB
bin 1
bin 2
bin 3
bin 4
bin 5
bin 6
bin 7
bin 8
bin 9
bin 10
downstream 10kB
downstream 20kB
Gene length

d.



e.

f.

Figure 5.1: Distribution of mononucleotide repeat motifs across the gene length.

a. Schematic representation of the strand asymmetry analysis, taking gene orientation into account. b. Enrichment in density of polyA and polyG motifs within genic regions. Error bars represent standard deviation from 1,000-fold bootstrapping with replacement across genes. c. Relative enrichment in density of polyN motifs at bins across the gene length. d. Non-template to template occurrences of polyN motifs across the gene length. e. Example of $polyA_7$ and $polyG_7$ non-template / template ratio at the different genomic bins across the gene length. f. Distance from the TSS and non-template / template polyN ratio for variable "N" length. Distance from the TES and non-template / template polyN ratio for variable "N" length.

5.3. Transcriptional strand asymmetry of indels at polyA motifs.

Having described the background frequency of polyN motifs and the associated strand asymmetries across transcribed regions, I investigated whether indels overlapping polyA and polyG motifs displayed asymmetries for non-template or template strands across 21 tumour types, controlling for differences in the background distribution of polyA and polyG motifs in the template and non-template strands.

It was found that across cancers, polyA motifs of length between 2-10bp were more mutable at the template strand; however, the levels varied by cancer type (Figure 5.2a). This was a surprising, novel result and implicates the influence of transcription on indel mutagenesis. What is unclear is whether the transcriptional strand bias is due to:

- an excess of DNA damage during transcription on the template strand, or
- an excess of transcription-coupled repair on the template strand for polyT tracts (that would result in an asymmetry of more indels on the non-template of polyT motifs which would be the equivalent to an excess of indels on the template strand for polyA motifs).

5.4. Mismatch repair deficiency enhances the transcriptional strand asymmetry of indels at polyA tracts.

To further explore these observations regarding indel asymmetries, the role of DNA repair processes was considered. In colorectal, uterus and stomach cancers, a deficiency of a repair pathway called Mismatch Repair (MMR) is often observed. These tumors have historically been reported to show a high level of instability at microsatellites (microsatellite instability, MSI). For these three cancer types, patient samples were separated into microsatellite stable (MSS) tumors which had proficient MMR, and MSI samples, which were deficient in MMR. Interestingly, the bias for indels on the template polyA motifs was accentuated in MSI samples (Figure 5.2c). It is already well-known that MMR plays an important role in reducing indel mutagenesis in general but this observation implicates MMR in the following ways:

- It could have a synergistic effect in concert with transcription-coupled repair usually acting more on the template strand.
- Or that there is simply much more in the way of transcription-related damage that the deficient MMR system is unable to cope with.

Regardless, the evidence indicates that deficiencies of the MMR complex contribute to the observed strand asymmetries for indels.

5.5. Transcriptional strand asymmetry of indels at polyG motifs.

For polyG motifs of length between 2-10bp, strong indel transcriptional strand asymmetry was not found in most cases, with the exception of indels in lung cancers, which exhibited strong preference for the non-template strand (Figure 5.2b). The strong asymmetry observed in lung cancers is likely to be a result of bulky adducts on guanines due to smoking-related carcinogens. These bulky adducts are more likely to be repaired on the transcribed strand by transcription-coupled repair, therefore accumulating more mutations on the non-transcribed strand. This has been reported experimentally (Chen et al. 1992), (Denissenko et al. 1998), (Nik-Zainal et al. 2015) and also through analyses of whole genome sequenced lung cancers in the past (Pleasance et al. 2010b). However, all of these reports have been for base substitutions and the effect of TCR on guanines at polyG tracts resulting in asymmetry of indels has not been previously reported.

To reinforce the observation, the level of transcriptional strand asymmetry relative as a function of gene expression was explored (Figure 5.2d). It was shown that increased levels of asymmetry occur at genes with higher expression levels, while genes that are not expressed or are expressed at low levels show minimal asymmetry (Figure 5.2d). This is in accordance with the fact that TC-NER activity is linked to gene expression levels and favours damage repairing at the transcribed strand and accumulation of more mutations at the non-transcribed strand.

Finally, transcription strand asymmetry levels for indels overlapping polyN motifs were explored at the different bins across the gene length. However, no conclusive patterns could be observed relative to the distance from the TSS or the distance from the TES, while the total number of indels was a limiting factor during this part of the analysis.

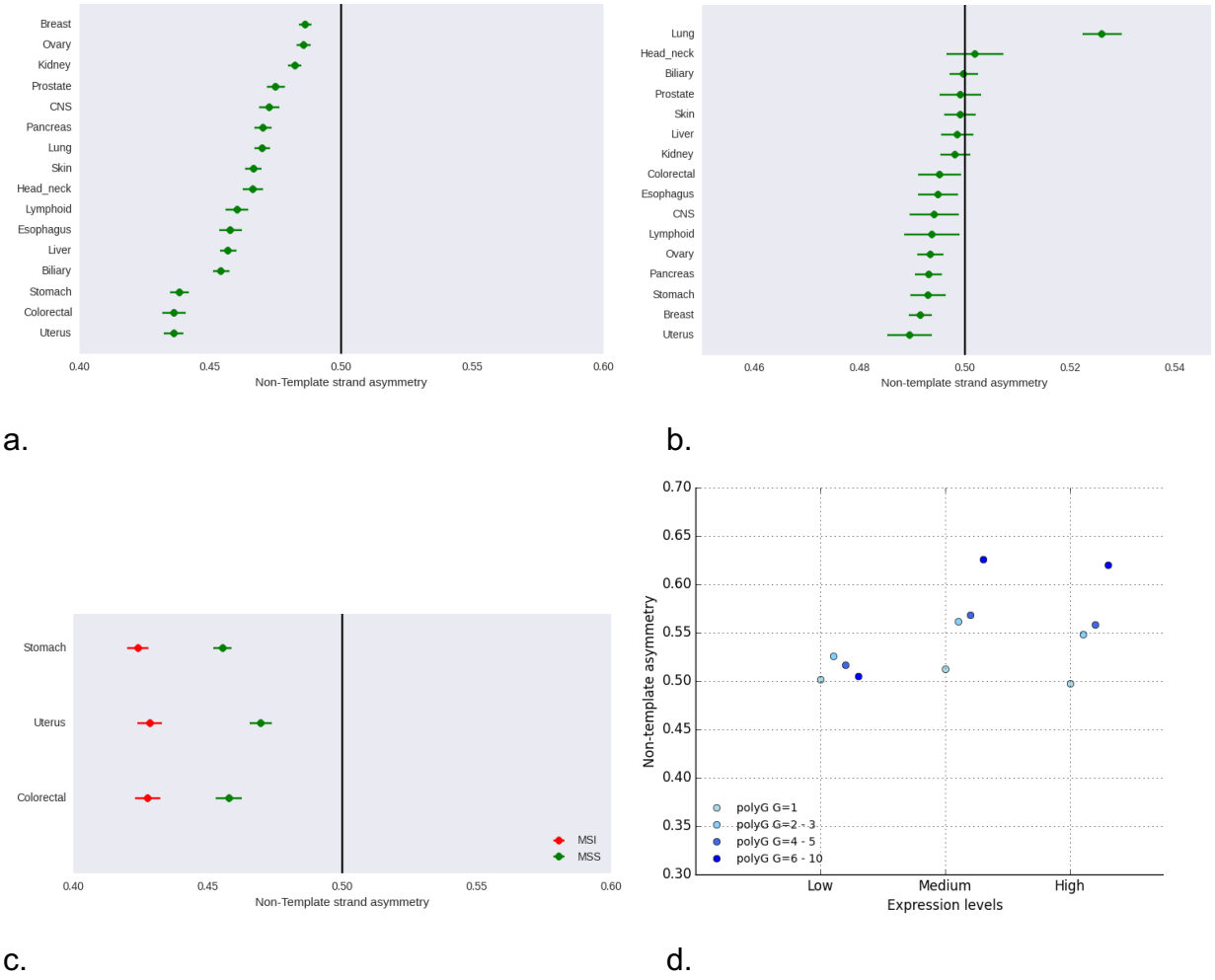


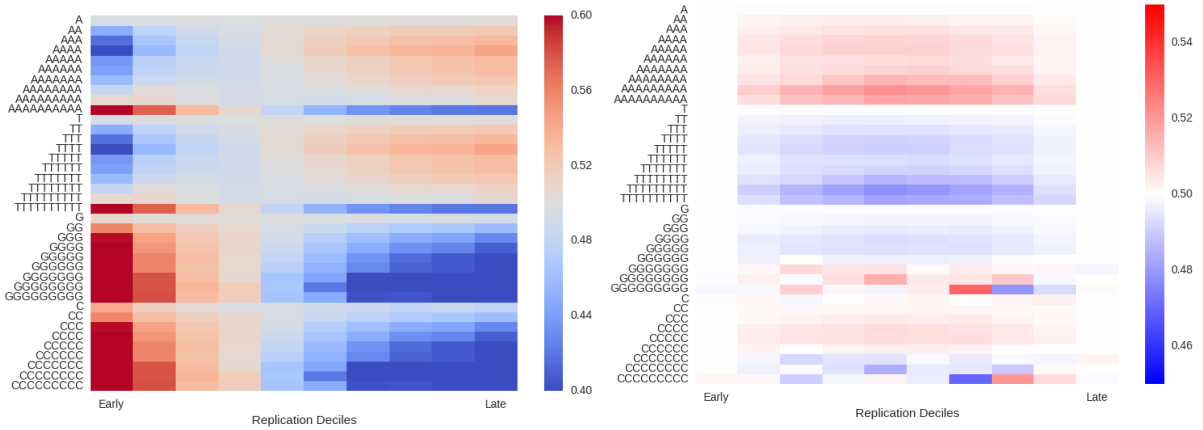
Figure 5.2: Indel transcriptional strand asymmetry across cancer genomes at polyN motifs.

Average indel transcriptional strand asymmetry at a. polyAs and b. polyGs is shown across genes with error bars indicating standard deviation from bootstrapping. Myeloid, cervix, bladder, bone and thyroid cancers were excluded due to small number of total indels (Table 4.1). c. Transcription-associated strand asymmetry at MSI and MSS samples for stomach, uterus and colorectal cancers for indels overlapping polyA motifs. d. Lung cancers show non-template preference for indel formation at polyG motifs which is associated with the expression of genes from a lung cell line and the length of the polyG tract.

5.6. Distribution of polyN motifs across replication deciles.

The distribution of polyN motifs was analysed across replication deciles. There was strong enrichment of polyG motifs in early-replicating genic regions, which is consistent with the fact that these regions tend to be gene-rich, while late replicating regions which are gene-poor were de-enriched in polyG motifs (Figure 5.3a). For polyA motifs, the enrichment showed the opposite pattern, with higher frequencies observed at late replicating, non-genic regions (Figure 5.3b).

The leading / lagging distribution of polyG and polyA motifs in the reference genome was subsequently also explored. PolyG distribution was relatively even across the replication timing deciles (Figure 5.3b). However, with the exception of the first and last deciles, polyA motifs displayed an enrichment for the leading strand, particularly for longer polyA motifs (Figure 5.3b).



a. b.

Figure 5.3: Distribution of polyN motifs across replication deciles.

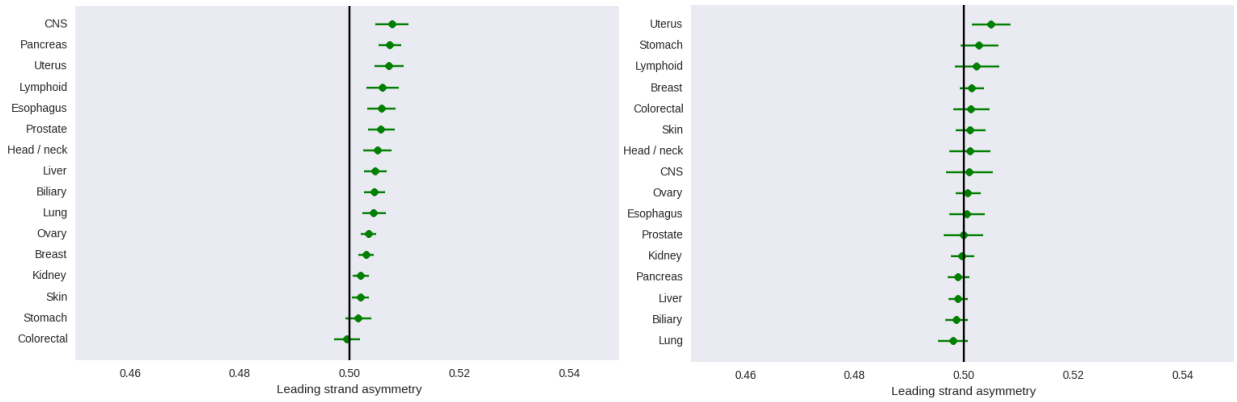
a. Relative enrichment of polyN motifs across the replication deciles. b. Leading versus lagging occurrences of polyN motifs across the gene length. Enrichment is calculated as occurrences at leading over occurrences at leading and lagging strand for each motif. Results from MCF-7 cell line Repli-seq data are shown. The enrichment at each decile was calculated as: Enrichment =

$\frac{density\ at\ decile}{density\ at\ decile + density\ across\ deciles}$. The leading to lagging ratio was calculated as: Ratio = $\frac{density\ at\ leading\ strand}{density\ at\ leading\ strand + density\ at\ lagging\ strand}$.

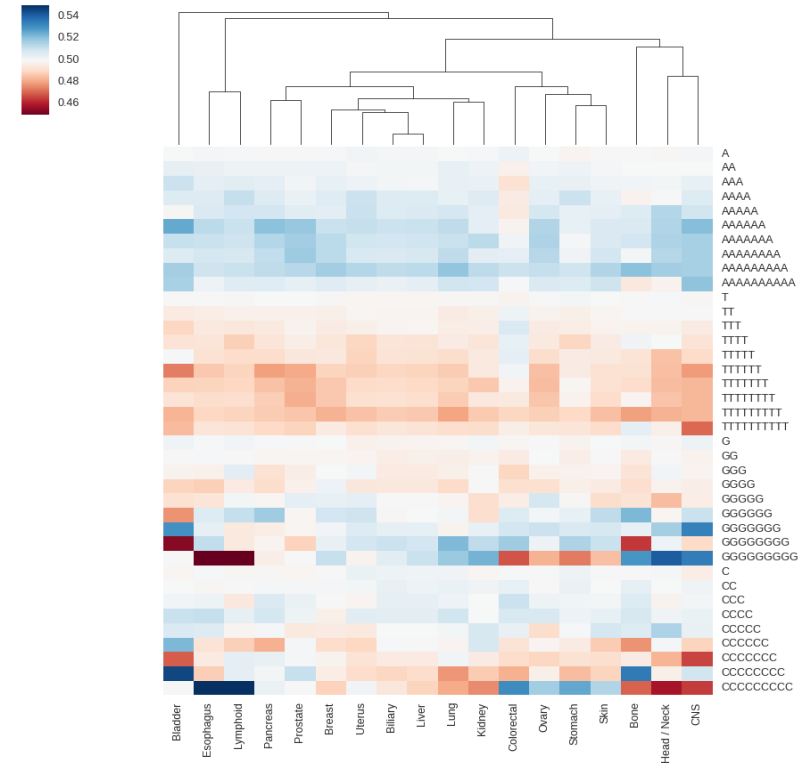
5.7. Replication strand asymmetry in mutability of polyN motifs in cancer genomes.

Next, it was investigated if indels display replication strand asymmetry at polyN motifs, with the analysis being performed across the different cancers (Table 4.1), corrected for the background distribution of polyN motifs across these domains. The analysis was restricted by the cell-of-origin of Repli-seq data, because for many tumour-types the corresponding cell-of-origin Repli-seq data was not available to make a direct comparison. As a result, the analysis was performed with Repli-Seq data from MCF-7 cell line for all tumour types. However, in chapter 3 it was shown that Repli-seq data of different cell lines are strongly correlated (r-squared>0.69 in all pairwise comparisons) and MCF-7 therefore serves as a good proxy to analyse multiple tumour types (Figure 3.2).

Replication strand asymmetry at polyA motifs was observed in multiple cancer types (Figure 5.4a,c). Nevertheless, the signal was weaker than that observed for the transcriptional strand asymmetry analysis. At polyG motifs the replication strand asymmetry levels observed were minimal (Figure 5.4b, c). Finally, the replication strand asymmetry at polyAs and polyGs was replicated in each decile separately, but no consistent patterns were observed across them. It was concluded that strong patterns of indel replication strand asymmetry at polyN motifs could not be observed.



a. b.



c.

Figure 5.4: Replication strand asymmetry at indels overlapping polyN motifs.

Indel replication strand asymmetry across cancer genomes at a. polyA motifs (left) and b. polyG motifs (right). Results from MCF-7 cell line Repli-seq data are shown. c. Hierarchical clustering of replication strand asymmetry by tissue type and polyN length. In all panels, ratio calculated as leading over lagging asymmetry of each polyN at each tissue.

5.8. Discussion.

The landscape of the distribution of mononucleotide repeat tracts was detailed across genic regions in the human genome. Unexpectedly, polyA repeat tracts were found to be enriched on the template strand in the reference genome and the bias was associated with the length of the repeat tract. However, it remains unclear which underlying mechanisms have disfavoured polyA repeats at the non-template strand. Future studies would need to explore this further to understand the origins of the observed bias. Nevertheless, this bias characterises transcribed regions of coding genes and it needs to be considered when exploring transcriptional strand asymmetries for indels overlapping mononucleotide repeat tracts.

The role of transcription-coupled repair has been previously explored for substitutions and their mutational signatures (Morganella et al. 2016), (Haradhvala et al. 2016). Here a novel method was presented to characterise indel mutations with respect to transcriptional strand asymmetry at genic regions (Figure 5.5). The observed levels of strand asymmetry are stronger at mononucleotide repeat tracts than those previously reported for substitution signatures (Morganella et al. 2016). These results also implicate TC-NER and MMR complexes as factors contributing to the observed strand biases at indels (Figure 5.2c-d). Although strong patterns for replication strand asymmetries were not observed, further analyses of other motifs or of novel indel signatures would be necessary to fully describe replication strand asymmetries at indels.

The described method serves as the first readout of systematic investigation of strand asymmetries for indels across cancer genomes. In addition, it was shown that mechanistic insight into mutational processes being operative in indel repair can be obtained from this method. Currently, the focus has been the investigation of polyN motifs; nevertheless, this approach could be expanded in the future to investigate other sequence motifs.

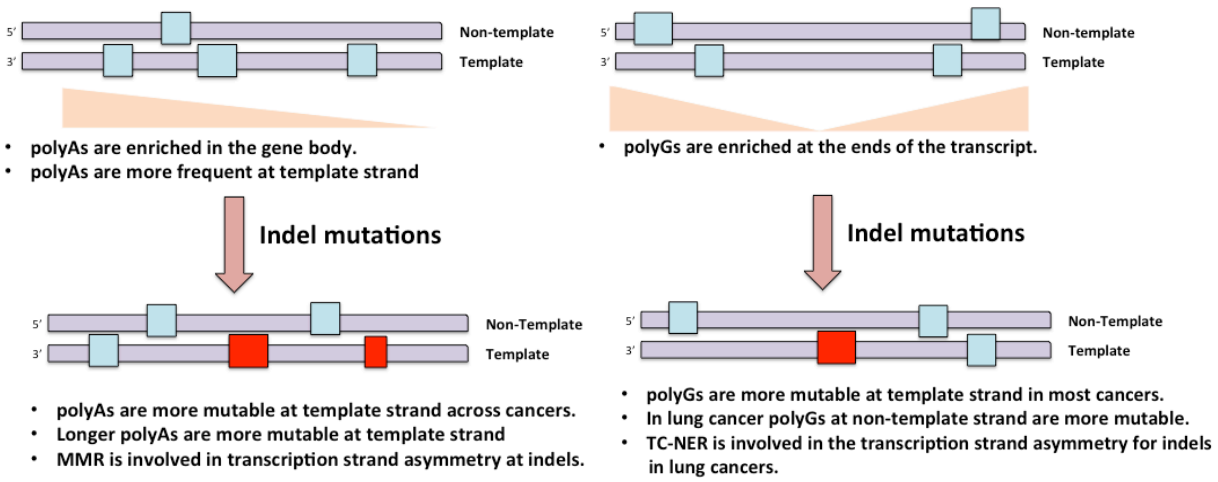


Figure 5.5: Schematic representation of transcriptional strand asymmetry for indels at mononucleotide repeat tracts.

PolyA motifs are enriched at the template strand and are more mutable at the template strand across cancers. MMR is involved in the observed transcription strand asymmetry for indels overlapping polyA motifs. PolyG motifs are enriched at the 5' and 3' end of genes and show non-template strand asymmetry in lung cancers. TC-NER is involved in the observed transcription strand asymmetry for indels overlapping polyG motifs in lung cancer.

CHAPTER SIX

6. Discussion and Future Work

The findings that were described in the earlier chapters of the thesis, bring forth multiple novel questions that remain to be answered. In this chapter these findings are put in perspective, while future work will be required to address the new hypotheses that stem and are described below.

6.1. Non-B DNA motif distribution in genomic sites.

Non-B DNA motifs display an inhomogeneous distribution along the human genome. Interestingly, they are enriched at a subset of regulatory regions, with most pronounced the enrichment at promoter regions, 5'UTRs and 3'UTRs (Figure 2.3a-c). Nucleotide resolution analysis of their distribution relative to functional elements indicated that they are not only enriched but are also positioned relative to those functional elements, including TSSs and TESs (Figure 2.4). Although it was shown that non-B DNA motifs are more mutable than their surrounding regions in a variety of tumour types, their distribution at functional sites in the human genome implies that they also have regulatory roles. This is further supported by a number of studies (Wittig et al. 1992), (Beaudoin and Perreault 2010), (Brooks and Hurley 2010), (Lam et al. 2013), (Quilez et al. 2016), (Gymrek et al. 2016), (Bay et al. 2017), (Armas et al. 2017), which have shown that non-B DNA motifs control expression levels and their disruption results in significant gene expression changes. Therefore, their mutational patterns in cancer genomes could have wider, functional implications that remain to be described; and as a result, their tumorigenic potential remains largely unexplored.

Moreover, the distribution of G-quadruplexes can be oriented relative to the transcriptional orientation (template / non-template) of a gene. It was found that there was transcriptional strand asymmetry around the TSS (Figure 2.4, Figure 2.5), which is in accordance with previous observations (Eddy et al. 2011). However, it was surprising to find that around the TES, G-quadruplexes are highly enriched at the non-template strand and are strongly depleted at the template strand (Figure 2.5). The non-template strand is the coding strand found at the mRNA level, and the enrichment might suggest that these structures have functional roles at the RNA level. However, further investigation would be required to understand their roles in relationship to transcription termination and transcript stabilisation and separate their putative effects at the DNA and RNA level. In particular, it would be of interest to investigate if disruptions of G-quadruplexes proximal to the TES, result in expression changes and whether the effect is specific to G-quadruplexes found at the non-template strand. Finally, putative intermolecular G-quadruplexes are abundant both at the TSS and the TES (Figure 2.5), raising the possibility that they are also implicated in transcriptional control.

6.2. Recurrent mutagenesis at non-B DNA motifs and potential functional consequences.

Non-B DNA motifs are more mutable than their surrounding environment. Their mutational enrichment is dependent on sequence characteristics of each non-B DNA motif and differs between their subcomponents. In particular, recurrent mutations are more likely to overlap non-B DNA motifs than non-recurrent mutations. A question that directly stems from these findings is whether recurrent mutations at non-B DNA motifs have functional and potentially tumorigenic effects. In support to that, as described earlier the distribution of non-B DNA motifs among functional elements is asymmetric (Figure 2.3a-c). For instance, promoters, which control expression levels, are enriched for multiple non-B DNA motifs. Therefore, it is plausible that a subset of recurrent mutations

at hypermutable non-B DNA motifs directly influence gene expression and have roles in cancer development.

Nevertheless, extensive analysis of an inverted repeat at the *PLEKHS1* promoter that is recurrently mutated in multiple patients across disparate cancer types did not find reproducible evidence for changes in its expression levels (Fredriksson et al. 2014). In addition, *PLEKHS1* is not a gene that has been traditionally implicated in cancer. This suggests that the recurrent mutations at the *PLEKHS1* inverted repeat, perhaps do not confer a selective advantage, or that a selective advantage has not been yet identified and characterised.

On the other hand, the *TERT* promoter harbours the most frequent recurrently mutated sites across non-coding regions, found in multiple cancer types and resulting in increased TERT expression levels (Weinhold et al. 2014), (Fredriksson et al. 2014). The functional effect of those mutations has been proposed to be the generation of ETS transcription factor binding sites (TFBS) that drive higher expression of the *TERT* gene (Huang et al. 2015). At the same region it has been shown that a G-quadruplex structure can form and can be stabilised by specific chemical compounds (Palumbo et al. 2009), (Lim et al. 2010). However, it was observed that the two most recurrently mutated sites overlap with that G-quadruplex motif (Figure 6.1c), (Weinhold et al. 2014), (Chaires et al. 2014). In particular, the two recurrent mutations overlapped with the G-runs and disrupted the potential for G-quadruplex formation. As a result, it is hypothesized that perhaps the functional effect of those mutations is not only the creation of novel TFBSs, but also the inactivation of a G-quadruplex structure, which has inhibitory transcriptional effects, that have been previously indicated (Palumbo et al. 2009), (Lim et al. 2010). Nevertheless, this has not been yet proven and it could be due to increased mutability at G-quadruplexes.

The described hypothesis is directly testable. For instance, expression levels of reporter assays using the *TERT* promoter as a regulatory element could be implemented. More specifically, mutations generating novel ETS TFBSs could be uncoupled from mutations

inactivating the G-quadruplex structure in those synthetic reporter assays. In addition, stabilisation of the G-quadruplex structure with chemical treatments, such as TMPyP4 or pyridostatin, in presence of novel ETS TFBSs, that do not inactivate the G-quadruplex, could be employed, to further explore the contribution of each in the regulation of the *TERT* promoter. In further support of the role of G-quadruplexes at promoters, a single nucleotide mutation at the *c-MYC* promoter, which disrupts the formation of a G-quadruplex that has transcriptional inhibitory roles (Figure 6.1b), results in ~3-fold increase in *c-MYC* gene expression (Siddiqui-Jain et al. 2002). On the other hand, a chemical compound that stabilises the same G-quadruplex structure results in a noticeable decrease of *c-MYC* expression levels (Siddiqui-Jain et al. 2002), suggesting that the G-quadruplex has a central role in modulating *c-MYC* expression levels. Finally, G-quadruplexes are observed more frequently in immortalised cell lines than in normal human cells, using immunofluorescence microscopy, immunohistochemistry and ChIP-seq, also suggesting potential roles in cancer (Hänsel-Hertsch et al. 2016), (Biffi et al. 2014).

Z-DNA motifs are enriched at promoters and in particular they have an enrichment peak relative to the TSS (Figure 2.3a-c, Figure 2.4a). Their regulatory effects in promoter regions have been described in a number of studies (Wittig et al. 1991), (Wölfl et al. 1996), (Liu et al. 2001). However, it remains unclear if their enrichment at recurrently mutated sites leads to changes in gene expression that could contribute to tumorigenesis. Moreover, polymorphic short tandem repeats account for approximately 10-15% of the variance in gene expression according to recent estimates (Gymrek et al. 2016), are causally implicated in many human disorders (Figure 1.5b), (Hannan 2010) and could account in part in the missing heritability problem of multiple complex diseases (Manolio et al. 2009). Therefore, it is plausible that a subset of recurrent mutations at tandem repeats could also have functional effects and be implicated in tumorigenesis. Furthermore, inverted repeats show an enrichment peak in relation to the TSS (Figure 2.4a); yet their roles at promoter regions have not been examined in detail until now.

The regulatory code of the human genome, through which transcriptional control is mediated, remains largely unknown. Additional work would be required to explore the regulatory effects of recurrent mutations at non-B DNA motifs. It remains unclear if the elevated mutational levels identified across cancers have functional consequences. Finally, if convincing examples are found, it would be of further interest to explore the frequency with which these events occur and quantify the magnitude of their effects.

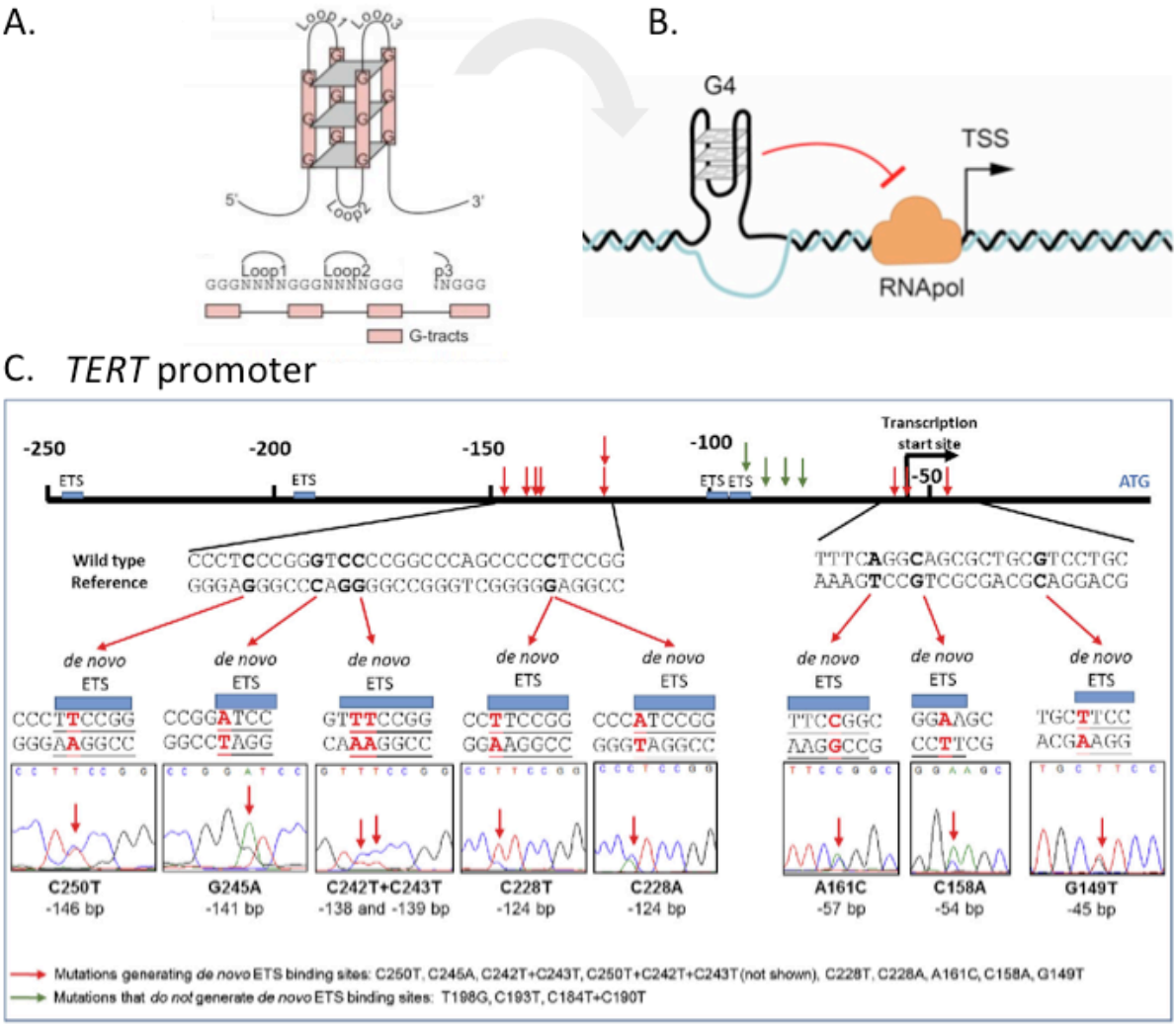


Figure 6.1: Inhibitory roles of G-quadruplexes upstream of the transcriptional start site.

a. G-quadruplex formation at the consensus motif. b. Schematic representation of the repressive roles of a G-quadruplex in transcription. Schematic from (Rhodes and Lipps 2015). c. TERT promoter recurrent mutagenesis generates novel ETS binding sites. Shematic from (Huang et al. 2015). However, the same recurrent mutations also inactivate the G-quadruplex. The two recurrent mutation sites were chr5: 1,295,228 and chr5: 1,295,250 (hg19) both of which had a C>T mutation

and both of which disrupted a G-quadruplex G-run and were found in 38 and 15 samples respectively (Weinhold et al. 2015).

6.3. Non-B DNA motif interactions with other players.

6.3.1. Nucleosome occupancy and positioning of non-B DNA motifs.

G-quadruplexes are found preferentially in nucleosome free regions (Hänsel-Hertsch et al. 2016). In addition, the nucleosome positioning relative to the position of mutations for certain substitution signatures and indel categories shows particular enrichment relationships (Morganella et al. 2016). Surprisingly, indels are not only enriched directly at G-quadruplexes but also at a distance of approximately 150bp away from them, at both sides of a 2kB window plot (Figure 2.6f) and is explained by the positioning of G-quadruplexes relative to nucleosomes (Figure 2.6g). This indicates that non-B DNA motifs should not be always viewed as isolated features of the genome, but in combination with other types of genomic elements. Z-DNA is another non-B DNA motif, which is inhibitory to the formation of nucleosomes (Garner and Felsenfeld 1987), (Wong et al. 2007). Therefore, a putative functionality of multiple non-B DNA motifs could be the displacement of nucleosomes at promoter regions. Finally, the association of indels with nucleosome positioning and non-B DNA motifs could involve unknown mutational mechanisms that could potentially be explored with novel indel mutational signatures.

6.3.2. Protein interactions with non-canonical secondary structures.

Evidence suggests that multiple transcription factors do not only recognise the primary sequence of the DNA, which encodes putative TFBSs, they also recognise the shape of the DNA molecule, where they bind. More specifically, DNA shape features such as the

helical twist and the minor groove width among others influence the binding of transcription factors at putative TFBSs (Mathelier et al. 2016). Interestingly, transcription factors that preferentially bind to non-B DNA have also been characterised. For instance, ADAR1 binds to Z-DNA over B-DNA (Herbert et al. 1995). Similarly, SP1 transcription factors bind with high affinity at G-quadruplex structures (Raiber et al. 2012).

Non-B DNA formation can be resolved by multiple enzymes, including helicases and topoisomerases. For instance, the RecQ helicase family members WRN and BLM have been found to resolve G-quadruplex and triplex structures (Sun et al. 1998), (Fry and Loeb 1999), (Brosh et al. 2001). Recently, DHX36 helicase was found to bind and unfold G-quadruplexes in the DNA and RNA level (Chen et al. 2018). Similarly, H-DNA is a substrate for a number of nucleases such as XPG, and FEN1 (Zhao et al. 2018), ribonuclease H1 processes R-loops (Lima et al. 2016) while the ERCC1-XPF complex is an endonuclease that cleaves cruciform structures and is associated with mutations that arise at inverted repeats (Lu et al. 2015).

Thus, non-B DNA formation and stabilisation is not independent of its interactions with various proteins, such as transcription factors, helicases, endonucleases and topoisomerases among many. Indicative of that, deletions due to MMR deficiency and HR deficiency, show different mutational enrichment patterns at disparate non-B DNA motifs (Figure 4.3e, Figure 4.4). It would be of further interest to investigate if in tumours that are deficient in helicases, endonucleases and topoisomerases, different categories of non-B DNA structures induce higher levels of genomic instability. Specifically, an investigation of the types of mutations that are most enriched at each non-B DNA category for tumours with such deficiencies would lead to a better understanding of the mechanisms underlying mutability at non-B DNA motifs.

6.4. Sequence characteristics of non-B DNA motifs influence their mutability.

For each non-B DNA motif category, biophysical properties, including spacer and arm length for inverted repeats, direct repeats and mirror repeats and loop size for G-quadruplexes influence their mutational patterns for substitutions and indels (Figures 3.3-3.4, Figure 4.4). In addition, mutability within non-B DNA motifs differs substantially at their subcomponents (Figures 3.3-3.4, Figure 4.4). As a result, statistical models of mutagenesis in cancer should consider the contribution of non-B DNA motifs in the heterogeneous landscape of mutagenesis. This in turn would permit the construction of more sensitive statistical models to investigate putative driver mutations.

As exemplified by inverted repeats at indels, the sequence characteristics of non-B DNA motifs and the underlying mutational processes contribute to differences in the likelihood of mutagenesis. Inverted repeats with spacer length of 5-6bps had a relative mutational enrichment that was more than 3.5-fold higher than that of other inverted repeats at deletions. In contrast, there was no noticeable mutational enrichment at insertions for inverted repeats of 5-6bps spacer length relative to inverted repeats with different spacer length (Figure 4.4a-b). Furthermore, the observed mutational enrichment was identified in microhomology-mediated deletions, but was not present in repeat-mediated deletions, implicating double strand break formation followed by aberrant repair (Figure 4.4c-d). Finally, microhomology-mediated deletions at inverted repeats were predominantly found at the arms but not at the spacer sequence, which was not the case for inverted repeats with different spacer length or for repeat-mediated deletions (Figure 4.4e-f). This in-depth analysis indicates that mutational patterns at non-B DNA motifs are influenced by the biophysical properties of each non-B DNA motif, differences in mutability of its subcomponents and the mutation type at play.

Sequence homologies at indel sites were also found at insertions and deletions. It was shown that the majority of large insertions result in the formation of “mini tandem repeats” and display strong sequence similarities between the inserted sequence and the insertion

site vicinity. Nevertheless, it remains unclear which is precisely the mechanism that is implicated in the formation of mini tandem repeats and it was suggested that it is most likely the result of replication slippage, in which case the same sequence is replicated twice. Additionally, it remains unknown if there are specific mutational processes that could increase the likelihood of “mini tandem repeat” formation.

6.5. polyA motif strand asymmetries across transcribed regions.

The biased distribution of polyA motifs at transcribed regions, with an excess frequency at the template over the non-template strand, as described in chapter 5 (Figure 5.1), to the best of my knowledge has not been described previously. The observed strand asymmetry of polyA motifs, which was exaggerated at longer polyA tracts, raises multiple questions that remain to be explored.

Firstly, it is unclear why the transcriptional strand asymmetries found at polyA motifs are not observed at polyG motifs. Secondly, it remains unknown what mechanisms have established such asymmetries for polyA motifs at transcribed regions in the human genome. More specifically, the observed transcriptional strand asymmetries could be the result of differences in DNA damage and repair at polyA and polyT motifs between the template and non-template strands or it could be the effect of yet unknown selective pressures. Thirdly, it is unknown if the observed asymmetry is specific to humans or is also found in other organisms. In particular, the question arises, are such asymmetries observed in distantly related species, which would imply that it is a general characteristic of transcription, or are they limited to humans and perhaps closely related species, that share the same repair pathways.

6.6. Mutational processes shape the indel landscape at transcribed regions.

The recent increase in the number of available whole genome sequenced (WGS) cancer patients, mediated primarily by two large consortia (ICGC & TCGA), provided a large dataset of high-quality indels, with a low false positive rate (Campbell et al. 2017). Therefore, a systematic characterisation of indels was feasible. In particular, a new method was developed to study transcriptional and replication strand asymmetries in indels that overlapped polyN motifs. This analysis unraveled unknown transcriptional strand asymmetries of indels, with indels overlapping polyAs exhibiting a preference for the template strand across tumour types (Figure 5.2a), whereas indels overlapping polyGs showed a preference for the non-template strand that was specific to lung cancers (Figure 5.2b). The observed transcriptional strand asymmetries of indels at polyN motifs could be attributed at least partly to TC-NER and MMR (Figure 5.2c-d), therefore providing mechanistic insight regarding the processes that are implicated.

However, multiple additional questions arise from this analysis. Firstly, are there indel strand asymmetries present in other motifs, different than the polyN motifs implemented here. This could be explored by a kmer motif analysis (e.g. across all pentamers) to find biases across a multitude of motifs and the mutational processes that are shaping those asymmetries. Additionally, indel signature extractions are currently being performed (Alexandrov et al. 2018), (Nik-Zainal, unpublished results). It would be of interest to examine if replication or transcription strand asymmetries are identifiable in any of these novel mutational signatures using these orthogonal methods.

6.7. Concluding remarks.

The aim of this thesis was to investigate the mutability of different non-B DNA motif categories across whole genome sequenced cancers. Firstly, the distribution of non-B

DNA motifs along the human genome was characterised and their enrichment at functional elements was demonstrated. Next, the mutational enrichment of each non-B DNA motif at substitutions, indels and rearrangements was investigated across multiple cancer types. It was found that non-B DNA motifs are more frequently mutated than their surrounding sequences for substitutions, insertions and deletions. However, for rearrangements the enrichment patterns at non-B DNA motifs were not conclusive in most cases, with the exception of rearrangements at inverted repeats. Models of mutagenesis along the genome were constructed separately for each type of mutation and it was indicated that non-B DNA motifs are predictive features for substitution and indel genome-wide mutagenesis.

Biophysical properties of non-B DNA motif categories were identified to be major determinants of differences in their likelihood of mutagenesis. Furthermore, the likelihood of mutagenesis was not homogeneous between the sub-components of each non-B DNA motif; exposed elements of non-B DNA motifs that were more likely to be found as single stranded DNA during structure formation, had an excess of mutations. Recurrent mutagenesis was also investigated in relationship to non-B DNA motifs. It was found that recurrent mutations are more likely to overlap non-B DNA motifs than non-recurrent, therefore obfuscating the identification of driver mutations in non-coding regions of the genome.

A characterisation of indels was performed across 2,575 whole genome sequenced cancers from 21 tissues. Non-B DNA motifs were enriched for indels and the level of enrichment was dependent on the non-B DNA motif and indel categories. Also, biophysical properties of different non-B DNA motifs such as spacer length for inverted, direct and mirror repeats influenced their relative mutational enrichment and could be coupled to specific DNA damage and repair processes. Sequence homologies at the site of indel formation indicated that the majority of large (>=10bp) insertions generate “mini tandem repeats” or in other words require the inserted or a highly similar sequence to already be present at the insertion site, whereas this was not the case for deletions.

In previous studies, transcription and replication strand asymmetries have been investigated for substitution mutational signatures (Morganella et al. 2016). However there is no available method to study such asymmetries at indels. Here, a novel method was developed and described in detail; this procedure allowed the investigation of transcription and replication strand asymmetries at indels using polyN motifs. It was subsequently shown that by applying this method mechanistic insight can be gained to understand the processes that are implicated in transcriptional strand asymmetries at indels. More specifically, mismatch repair and transcription-coupled nucleotide excision repair were found to contribute to the observed transcriptional strand asymmetries at indels, which until now was not known.

CHAPTER SEVEN

7. Materials and Methods.

7.1. Somatic variants from cancer data (Chapters 2-3).

The results presented in chapters 2-3 preceded those on the following chapters and thus contained fewer cancer types and patients, which were available at the time (Table 2.1). For chapters 2-4, data were obtained from whole genome sequenced (WGS) cancers across multiple cancer types including breast cancers from (Nik-Zainal et al. 2016) and from nine other cancer types publicly available in ICGC (ICGC and TCGA projects), (Table 2.1). The ICGC project codes were used to describe each of the cancer types across chapters 2-4. The reference assembly GRCh37 (hg19) of the human genome was used throughout the work.

In particular, a total of 1,809 WGS cancers were analysed. These included 560 breast cancers (Nik-Zainal et al. 2016), 242 pancreatic cancers (148 PACA-CA and 94 PACA-AU), (Waddell e al. 2015), (Notta et al. 2016), 72 ovarian cancers (OV-AU), (Patch et al. 2015), 264 liver cancers (LIRI-JP), (Fujimoto et al. 2016), 120 prostate cancers (PRAD-CA), (Fraser et al. 2017), 98 esophageal carcinoma cancers (ESAD-UK), 74 renal cell cancers (RECA-EU), 239 pedriatic brain cancers (PBCA-DE), 100 malignant lymphomas (MALY-DE) and 40 gastric cancers (GACA-CN), (ICGC and TCGA projects). Across every tumour and matched normal sample sequencing coverage exceeded 25X.

Somatic variant calling for substitutions, insertions, deletions and structural variations was performed using the Wellcome Trust Sanger Institute Cancer Genome Project WGS pipeline as previously described in (Nik-Zainal et al. 2016). The pipeline includes an Expectation-Maximisation-based algorithm to identify substitutions (CaVEMan), (Nik-Zainal et al. 2012), (Jones et al. 2016), Pindel, which is an insertion and deletion

deletection algorithm (Ye et al. 2009) and a structural variant detection algorithm that is based on de Brujin graphing for the identification of somatic rearrangements and local assembly for mapping breakpoints to base pair level.

Mutation simulations for 10% of substitutions in each cancer type were performed controlling for trinucleotide context of each mutation and proximity (within a 50kB window from the original mutation).

7.2. Somatic variants from cancer data (Chapters 4-5).

For chapters 4-5, a larger cohort of cancer samples was available during the analysis. Data were obtained for WGS cancers from ICGC under the project PanCancer Analysis of Whole Genomes (PCAWG). The patient cohort included 46 cancer projects from 21 organs. In total, 2,575 WGS patients were analysed using the GRCh37 (hg19) reference assembly of the human genome.

Somatic indel calls were performed using three pipelines from four somatic variant callers. These were the Wellcome Sanger Institute pipeline, the DKFZ/ EMBL pipeline and the Broad Institute pipeline and indels were identified as described in (Campbell et al. 2017), with somatic variant false discovery rate of 2.5%. Indel calling was performed by those algorithms and only indels called by at least two of the callers were analysed (Campbell et al. 2017), therefore generating a conservative dataset (Table 4.1). As a result, the false negative rate of indel detection could be higher than that of other methods, and of each pipeline separately, which implies that many indels present in the samples were not identified successfully. However, because of the large number of WGS tumour samples available, a sufficient number of indels remained (Table 4.1). Finally, for a small subset of indels, the indel calls were visually examined using JBrowse Genome Browser (Buels et al. 2016), to inspect the number of reads reporting the indel, if the indel calls were biased towards the end of the sequencing reads or if there were other systematic biases between the normal and tumour sequencing reads; such biases could not be identified.

The distance between each pair of consecutive indels was calculated for each of the patients. Indels in different chromosomes were excluded, because their pairwise distance could not be defined. Patients without multiple indels in the same chromosome were also excluded. The same analysis was also performed separately for insertions and deletions to generate Figure 4.1a. The distribution of insertions and deletions by size across cancer types was measured as well as the ratio between them across patients (Figure 4.1b-d)

Deletions were classified into microhomology-mediated and repeat-mediated deletions as described in (Nik-Zainal et al., 2016). The two types of deletions differ in that repeat-mediated deletions occur at repeat sequences, whereas microhomology-mediated deletion sites display microhomology between the deleted sequence and the 3' region directly downstream and are thought to be caused by aberrant repairing of double-strand breaks. Mutational density analysis of indels and of repeat-mediated and microhomology-mediated deletions across the non-B DNA motif categories was performed (Figure 4.3d-e).

7.3. Reference non-B DNA annotations.

Genome-wide maps for each non-B DNA motif category for the human reference genome (hg19) were derived from (Cer et al. 2013). The analysis of non-B DNA motifs was performed for seven categories of non-B DNA motifs. These were inverted repeats, direct repeats, mirror repeats, short tandem repeats, G-quadruplexes, H-DNA and Z-DNA; the definitions of which are found below.

- A mirror repeat is a sequence reading the same both in the 5' and 3' direction, with a spacer sequence that does not show the homology pattern. It has length of at least 20nt and arm size of at least 10nt.
- A subset of mirror repeats, which can be identified from the primary nucleotide sequence of the human genome, can form Hinged DNA (H-DNA). H-DNA is a triple

helical structure connected through Hoogsteen bonds. H-DNA sequences have >90% AG content, arm length of >=10nt and spacer size <8nt.

- Z-DNA is a left-handed double helical structure with a characteristical zigzag pattern. It can form at alternating purine-pyrimidine tracts >=12nt (excluding AT repeats).
- Direct repeats consist of two repeats of the same sequence, interspersed by a spacer sequence. Here they have an arm length of >=10nt and maximum size of 300nt.
- Short tandem repeats also known as microsatellites are defined as motifs of 1-9nt, repeated at least 3 times with a minimm length of 9nt.
- Inverted repeats consist of a sequence followed by its reverse complement. They have arm length >=6nt, spacer size up <=100nt and can fold in hairpin or cruciform structures.
- G-quadruplexes are non-canonical nucleic acid structures held together with Hoogsteen bonds. Here they were defined as four or more G-runs of at least 3 guanines that are separated by spacer sequences of 1-7nt.

The distribution of non-B DNA motifs relative to genic sites, including transcription start sites, transcription end sites, translation start sites and translation end sites was investigated. The Ensembl reference genic annotation was used (hg19) and window plots were generated centered at the genic site of interest and measuring the distribution of non-B DNA motif categories relative to it (Figures 2.4a-d).

A script based on motif finding using regular expression patterns was developed in python to generate genome-wide maps of G-runs in the human genome. Pairs of G-runs were interspersed with variable regions of 1-7bp. In total, one to four consecutive G-runs were searched for, the later of which denoted the consensus G-quadruplex motif. These maps were used to investigate if the pattern of enrichment at functional sites differed incrementally between the number of G-runs a motif had and to investigate putative strand asymmetries by orienting them relative to the transcription direction of genes (Figure 2.5a-d).

7.4. Genomic element partitions and chromatin states.

The Ensembl Regulatory Build is a genome-wide annotation map of regions involved in gene regulation (Zerbino et al. 2015). The annotated categories include diverse regulatory elements, namely promoters, promoter flanking regions, enhancers, CTCF binding sites, transcription factor binding sites and open chromatin regions. The density of each non-B DNA motif was calculated at each of the annotated Ensembl Regulatory Build categories and was compared to the density across all the catogies to calculate the enrichment at each element involved in gene regulation (Figure 2.2a). Clustering and figure plotting were performed with the “seaborn” package in python.

The Segway algorithm generates chromatin states which are partitions of the genome with genomic annotation labels that are derived using an unsupervised pattern discovery algorithm with inputted chromatin modification data. In chapter 2, chromatin states were defined with Segway as described in (Hoffman et al. 2012), (Hoffman et al. 2013) using chromatin modification maps previously generated with data from (The ENCODE Project Consortium 2012) for six human cell lines (GM12878, H1-Hesc, HepG2, HUVEC, K562, HelaS3). The labelled partitions were namely: CTCF, DNase, transcription associated, candidate strong enhancer, candidate weak enhancer, low activity proximal to active states, promoter flanking, active promoter, inactive promoter, heterochromatin-repetitive-copy number variation and Polycomb repressed.

The enrichment patterns of non-B DNA motifs across the different chromatin states were investigated. The density of each non-B DNA motif category at each chromatin state was compared to that across all chromatin states at each cell line from which the mean enrichment across the six human cell lines was calculated. Hierrarchical clusering of chromatin states and plotting was performed with the python package “seaborn” using default parameters (Figure 2.2b).

7.5. Epigenomic data.

Narrowpeak files for DNase-seq and sevaral histone modifications (H3K4me1, H3K4me3, H3K427ac, H3K36me3, H3K9me3, H3K27me3) were derived from (Roadmap Epigenomics Consortium 2015) and BAM files for DNA-seq and for the same histone modification alterations were derived from (The ENCODE Project Consortium 2012). In particular, HMEC narrowpeak files were used to model breast cancer. Similarly, PANC1, HepG2 and GM12878 cell line narrowpeak files were used to model pancreatic cancer, liver cancer and malignant lymphoma respectively, while for modelling ovarian cancer, esophageal cancer, pedriatic brain tumour, gastric cancer and renal cell cancer narrowpeak files from ovary, esophagus, fetal female brain, stomach mucosa and fetal kidney primary tissue were used accordingly. To validate the findings using narrowpeak files, BAM files for DNA-seq and the histone modifications were downloaded from (The ENCODE Project Consortium 2012) and analysed using MCF-7 cell line to model breast cancer.

7.6. Repli-Seq data.

Repli-seq data were obtained from (The ENCODE Project Consortium 2012) and processed as described in (Morganella et al. 2016) for multiple cell lines to invstigate the relationship between mutability and replication timing. More specifically, reference coordinates for replication timing were derived for 14 different cell lines. These were namely IMR90, GM12801, HUVEC, BJ, NHEK, GM12813, GM12812, MCF-7, GM06990, HeLa-S3, BG02ES, HepG2, GM12878 and K562. Repli-Seq data for MCF-7 cell line were used to model the relationship with mutability in breast cancer and similarly HepG2 cell line Repli-Seq data were used for liver cancer, GM12878 cell line Repli-Seq data were used for malignant lymphoma and MCF-7 cell line Repli-Seq data were used for all other cancer types available. Replication timing was measured at genomic intervals of the human genome using the command “bedtools map”. The correlation for replication timing

between any two cell types across 500kB bins of the human genome was calculated and is shown in (Figure 3.1b).

7.7. Genome-wide models of mutability based on epigenetics, replication time domains and non-B DNA motifs.

Partioning of the human genome in 500kB bins was performed. Centromeric regions, simple, low complexity regions and regions of excessive sequencing depth (UCSC Top 0.01 Hi Seq Depth) were used to identify bins of low mappability using the command “bedtools coverage”. Additionally, the first and last bin from each chromosome of the human genome and the sex chromosomes were excluded as well as any bin where <50% of the bases were mappable or where replication timing data is missing. This resulted in 5,581 genomic bins of 500kB size. All quantiles except for the replication timing were transformed as x’=log2(1+x).

GC content and nucleotide composition at each interval was calculated with the command “bedtools nuc”. To map reads from DNA-seq and histone modification BAM files from (The ENCODE Project Consortium 2012) at each genomic interval, “bedtools multicov” was used. To calculate the number of non-B DNA motifs, mutations, genomic features and narrowpeak files from (Roadmap Epigenomics Consortium 2015) at each genomic segment, BEDtools “intersect” function was used. A principal component analysis was performed (Fig. 3.2a) using the R command “princomp”.

Partial correlations were calculated in R with the package “ppcor” (Kim 2015). Partial correlations were applied to measure the relationship between mutations and non-B DNA motifs, controlling for the effect of epigenetic markers and replication timing, in 500kB genomic bins (Fig. 3.1e).

Linear and random forest regression.

The relationship between the number of substitutions, indels and rearrangements at each 500kB genomic bin and multiple genomic features, including non-B DNA motifs, histone modifications and replication timing was investigated systematically. Two predictive models were constructed in R; the first model was based on linear regression and the R command “lm”, whereas the second model was based on random forest regression using the R package “randomForest” (Figure 3.2b). For both models, 10-fold cross-validation was used, and each model was trained using 90% of the data and tested using the other 10%. For random forest regression, the variable importance was calculated using permutation testing with the “pRF” package in R (“n.perms = 200”, “mtry = 4”), (Figure 3.2c-d). The two models were applied independently to each cancer type; prostate cancer was excluded because cell of origin epigenetic data were not available.

7.8. Analysis of mutagenesis at non-B DNA motifs.

For each 500kB genomic window, its size was denoted as B, while the size of the window covered by a non-B DNA motif was termed b. Mutations were separated in those overlapping the non-B DNA motif, which were termed m and those that did not overlap the non-B DNA motif, termed n. The mutational enrichment at each non-B DNA motif was calculated as:

$$r=m(B-b)/nb,$$

with genomic bins with b=0 excluded from the analysis.

The expected values and variances in calculated ratios were adjusted to account for correlations using the following equations:

$$E[X/Y] = E[X]/E[Y] - Cov[X, Y]/E[Y]^2 + Var[Y]E[X]/E[Y]^3$$

$$Var[X/Y] = (E[X]/E[Y])(Var[X]/E[X]^2 - 2Cov[X,Y]/E[X]E[Y] + Var[Y]/E[Y]^2)$$

Inverted, direct and mirror repeats were studied in relationship to mutagenesis for different spacer and arm lengths. The mutational density at spacers and arms was calculated for each motif and averaged across all occurrences of the motif category. Subsequently the mutational density of spacers and arms was plotted in relationship to spacer and arm length for substitutions (Figure 3.3a, Figure 3.4) and was also corrected for trinucleotide context of substitutions (Figure 3.3b).

The G-quadruplex motif subcomponents were the G-runs and the interspersed loops. The mutational density was calculated separately in each sub-component across the motif occurrences and the fraction of mutational densities was calclulated (Figure 3.3c). In addition, the trinucleotide context of substitutions was considered in Figure 3.3d. G-quadruplex motifs were further separated based on the average length of their interspersed loops into two groups (less or equal than 3nt or longer than 3nt) and the mutational density was compared between the two groups (Figure 3.3e). Bootstrapping with replacement was performed to calculate the standard errors.

Analysis of the distribution of non-B DNA motifs at mutation sites with single nucleotide resolution was performed using 2kB window plots, centered at substitution or indel mutations. The enrichment of non-B DNA motifs at each position was calculated from dividing the number of occurrences of a non-B DNA motif category at a position over the median number of its occurrences across the window (Figure 2.6c-f). Micrococcal nuclease sequencing (MNase-seq) data for K562 cell line were derived from (The ENCODE Project Consortium 2012) to investigate the relationship between G-quadruplexes, nucleosome positioning and mutability (Figure 2.6f-g). A heatmap plot of nucleosome occupancy was produced centered at G-quadruplexes and measuring the MNase-seq score across the window using deepTools (Ramírez et al. 2014).

The functions “intersect” and “coverage” were used to find the overlap of indels and non-B DNA motifs, as well as the presence or absence of non-B DNA motifs at indel sites. For the non-B DNA motif analysis, indel controls were same size with the indels and shifted

500bp away randomly to the left or to the right of the indel site (Figure 4.3). The association between insertion and deletion mutational density at inverted, direct and mirror repeats was investigated as a function of spacer length for spacer lengths of 0 to 10nt (Figure 4.4a-d). In addition, mutational density of indels was compared at spacers and arms for inverted repeats of different spacer lengths for indel mutation categories (Figure 4.4e-f).

7.9. Recurrent mutagenesis in cancer genomes.

At each genomic site in the genome, mutational recurrency across patients in each cancer type was calculated separately for substitutions and indels. The script to identify recurrent mutations is provided in (Georgakopoulos-Soares et al. 2018). A truncated Poisson model was constructed in R with “mle” function using “stats4” R package (Figure 3.5a). The command “bedtools intersect” was used to calculate the frequency of overlap of recurrent and non-recurrent mutations for each non-B DNA motif category (Figure 3.5b-c) and Mann Whitney U testing was performed to measure statistical significance.

7.10. Template / Non-template strand asymmetries at the reference human genome.

Gene annotation from Ensembl was followed (Aken et al. 2016). GC-skew is a measure of bias in Gs or Cs in template and non-template strands. GC-skew was calculated as (G-C) / (G+C) for windows of 100 bp around the TSS and TES. Similarly, AT-skew was calculated as (A-T) / (A+T) for windows of 100 bp around the TSS and TES.

Genes in the positive and negative orientations were separated to determine the direction of gene transcription. Scripts were written in python to identify non-overlapping polyN motifs of size 1-10bp and orient them in terms of transcription direction at genic regions.

Template motifs were the motifs in: i) positive gene orientation and negative genome strand, ii) negative gene orientation and positive (reference) genome strand. Non-template motifs were the motifs in: i) positive gene orientation and positive genome strand (reference), ii) negative gene orientation and negative genome strand. Bedtools intersect command was implemented to calculate motif occurrences in template and non-template strands across genic regions.

To investigate the effect of the distance from the TSS and the TES across the gene length, for genes with unequal gene length, each gene was divided in ten genomic bins of equal size. Also, two additional bins upstream from the TSS and two bins downstream of the TES, each 10kB in size, were added. Then, the frequency and the strand asymmetry bias of polyN motifs was calculated in each genic bin (Figure 5.1b-e).

Strand asymmetry bias was calculated as:
(motif occurrences at non-template strand) / (motif occurrences at template strand)

The distribution of polyN motifs at the template and non-template strands relative to the TSS and the TES were calculated using bedtools intersect command using the gene orientation approach described earlier to generate (Figure 5.1b-e). Bootstrapping using random sampling across genes with replacement for equal sampling experiments to the number of genes was performed from which the standard deviation of the strand asymmetry bias was calculated.

7.11. Template / Non-template strand asymmetries in cancer.

The number of indels overlapping motifs found in the template or non-template strands were calculated using the bedtools intersect command. Enrichment was calculated for the vector of genes, reporting the number of polyN motif occurrences and the number of overlapping motifs as:

A = (indels overlapping motif at non-template) /(motif occurrences at non-template)
B = (indels overlapping motif at template) /(motif occurrences at template)
Enrichment = A / (A+B)
with motifs representing polyN tracts, at genic regions (Figure 5.2).

Bootstrapping with replacement was performed, randomly selecting the indels overlapping motifs at template and non-template strands from each randomly selected gene, for equal sampling experiments to number of genes, from which the standard deviation of enrichment was calculated.

Mismatch repair deficient samples were identified using genome plots and mutational signature profiles of each patient for stomach, uterus and colorectal tumours. Subsequently, indel strand asymmetry levels were compared between MSS and MSI samples to investigate the role of mismatch repair in transcriptional strand asymmetries (Figure 5.2c).

7.12. Replication timing strand asymmetries at the reference human genome.

The frequency of polyN motifs was investigated across replication deciles using MCF-7 cell line Repli-Seq data. The enrichment at each decile was calculated as:

Enrichment = (density at decile) / (density at decile + density across deciles).

Similarly, the frequency of polyN motifs was calculated at leading and lagging replicative strands, which were inferred as described in (Morganella et al. 2016), using the reference human genome strand directionality. The ratio for the leading to lagging asymmetry was calculated as:

Ratio = (density at leading strand) / (density at leading strand + density at lagging strand).

7.13. Leading / Lagging strand asymmetries in cancer.

The number of indels overlapping motifs found in the leading or lagging strands were calculated using the bedtools intersect command and enrichment was calculated from the number of polyN motif occurrences and the number of overlapping motifs as:

A = (indels overlapping motif at leading strand) /(motif occurrences at leading strand)
B = (indels overlapping motif at lagging strand) /(motif occurrences at lagging strand)
Enrichment = A / (A+B)
with motifs representing polyN tracts (Figure 5.2)

7.14. RNA-seq and transcriptional strand asymmetry at polyN motifs for indels.

For the comparative analysis between expression levels and transcriptional strand asymmetry in lung cancers, the cell of origin cell line IMR-90 was used from Roadmap epigenomics project (Roadmap Epigenomics Consortium et al. 2015). PolyG tracts were grouped according to their length to investigate if the length of polyG tracts was associated with transcriptional strand asymmetry at indels across gene expression quantiles (Figure 5.2d). In particular, genes were grouped in three expression level quantiles, namely “low”, “medium” and “high” based on the associated RPKM gene expression values.

7.15. Sequence similarities at indel sites.

Indels >=10bp were examined for identical copies of the sequence that was inserted or deleted in the vicinity of the insertion or deletion site (flanking 500 bps on each side),

(Figure 4.5a) and the enrichment was calculated from comparing the frequency at the indel controls. Similarly, it was also investigated if the inverted or mirror sequence of those indels could be identified in the indel vicinity more frequently than in the control indels (Figure 4.5b-c).

Hamming distance is a similarity metric measuring the number of positions that differ between two sequences, with zero distance representing identical sequences. Scripts were written in python to measure the similarity between insertions / deletions of: i) at least 5 bp length, ii) at least 10bp length and the indel region using the hamming distance measure. The search space was the indel region, defined as +/-500bp from the indel site. Two controls were designed, the first shifting the indel site by 3kB and the second scrambling the reference sequence at the indel site, therefore controlling for position in the genome and nucleotide composition. To measure the hamming distance of the mirror / inverted sequence of an indel, its mirror sequence or inverted sequence respectively was used (Figure 4.5d-f).

7.16. Enrichment of indel categories at regulatory elements.

Transcription factor binding sites, promoter regions and open chromatin regions were derived from the Ensembl Regulatory Build (Zerbino et al. 2015). The density of insertions and deletions as well as repeat-mediated deletions and microhomology-mediated deletions was calculated. Enrichment was calculated as the density of an indel category at a position relative to the site of the regulatory element over the median density of an indel category across a 2kB window relative to the regulatory element (Figure 4.6a-c).

7.17. Motif analysis using all kmers of 1-7 nucleotides length.

A motif finding algorithm using regular expressions was developed to scan sequences for motif occurrences both at a list of DNA sequences of interest and their reverse

complement. Indels were separated by indel category into insertions and deletions and by tissue of origin. The motif search was performed separately for insertions and deletions at a local window (+/-150bp), and separately for each tissue of origin. There were 2,470,494 indels in total and 21,844 kmer motifs, resulting from all possible kmers of length 1-7nt. The number of motif occurrences in each case was corrected by the total number of nucleotides searched in. A binomial test was performed, in which the expected proportion was derived from the ratio of the number of nucleotides scanned in insertion sites over the number of nucleotides scanned in deletion sites. Bonferroni correction was performed correcting for multiple testing regarding the 21,844 kmer motifs used. To identify motifs that were differentially enriched across organs, a threshold of 18 out of 21 organs was used with a p-value < 0.01 across each of those tissues. The python package seaborn was used for clustering and figure generation with the function “clustermap” (Figure 4.2a).

CHAPTER EIGHT

8. Appendix

8.1. Introduction: Massively parallel reporter assays (MPRAs).

The regulation of gene expression is a set of processes which together guarantee that each cell in the human body produces the necessary genes for its functionality. As a consequence, aberrant regulation of transcription results in a number of disorders (Lee et al. 2013). In support to that, the majority of single nucleotide polymorphisms (SNPs) that are associated with disease are found in non-coding, regulatory regions of the genome (Maurano et al. 2012). DNA regulatory elements, including promoters and enhancers, contain short sequence motifs, also known as transcription factor binding sites (TFBSs), at which transcription factors can bind to modulate gene expression levels. Indicative of their importance, expression of the four Yamanaka transcription factors is sufficient to induce pluripotency (Takahashi and Yamanaka 2006). However, the gene regulatory code remains poorly understood to date.

Sequencing technologies are advancing at an unprecedented pace and in combination with the rapid decline in the associated DNA synthesis and sequencing costs, parallel high-throughput experiments are becoming affordable. Among them, a multitude of high-throughput reporter assay methods have been invented to investigate systematically the roles of regulatory elements (Inoue and Ahituv 2015). In particular, massively parallel reporter assay (MPRA) experiments implement the recent technological advances to assay tens of thousands of regulatory sequences in a single experiment (Melnikov et al. 2012), (Levo and Segal 2014), (Figure 8.1).

The potential of MPRA technologies is clear; they can be implemented to examine existing or synthetic sequences to increase our understanding of the regulatory code and examine disease variants. For instance, they can be used to examine the role of SNPs and expression quantitative trait loci (eQTLs) found at regulatory regions, which could modulate expression levels. In addition, they can be used to examine the promoter or enhancer potential of a putative sequence or to investigate how combinations of TFBSs influence expression levels. In addition, synthetic sequences not present normally in nature can also be designed, which becomes useful when testing a hypothesis. For instance, the role of homotypic and heterotypic TFBS clusters and the relative distances of TFBSs can be investigated using saturation experiments. Nonetheless, designing in parallel thousands of MPRA sequences remains difficult.

There are currently no available bioinformatic methods that could potentiate the design of MPRA experiments by non-experts. Additionally, methods allowing fast and accurate design of MPRA sequences would increase the popularity of high throughput reporter assay experiments in the community.

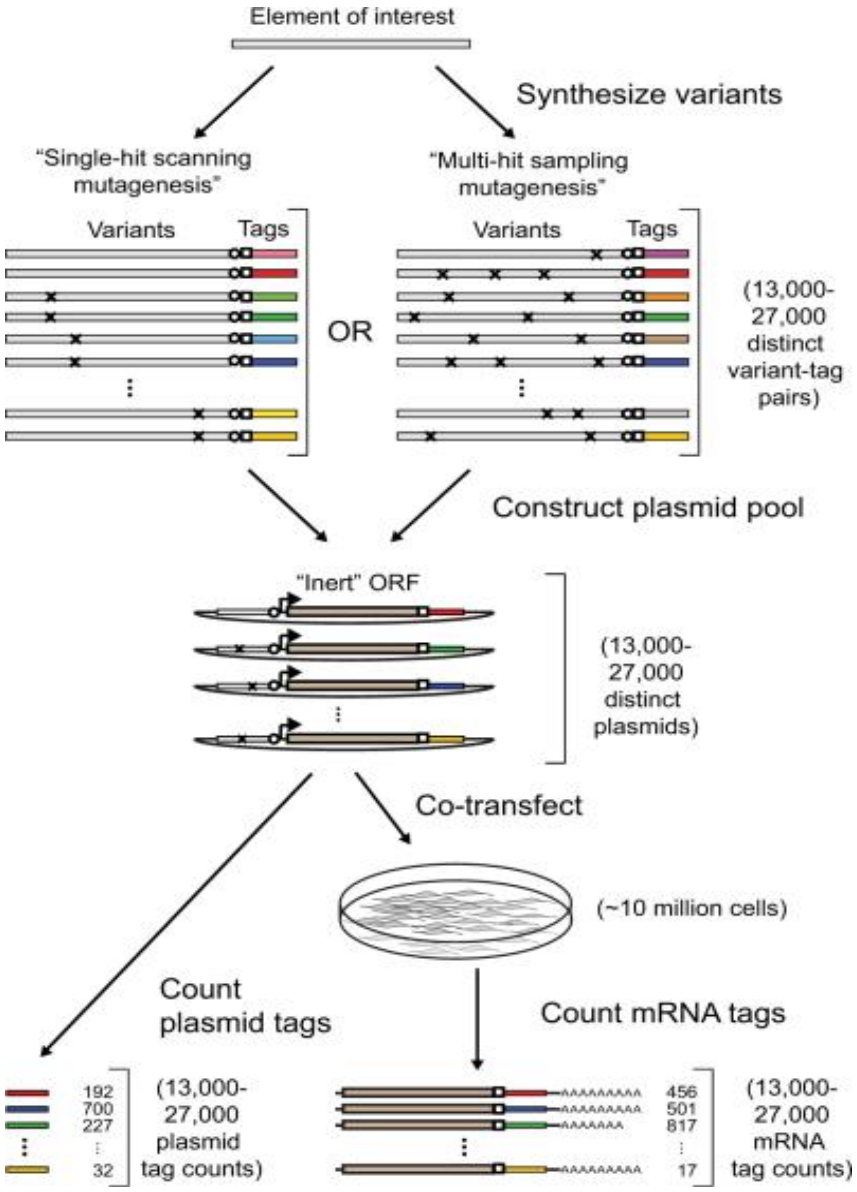


Figure 8.1: Massively parallel reporter assay experimental design.

Oligonucleotides with variants of interest are synthesized and barcoded with unique tags. In a next step, the oligonucleotides are PCR amplified and then inserted in reporter vectors. By counting the number of occurrences of the barcode tags at the DNA and mRNA level, the expression levels of each variant can be inferred. Schematic from (Melnikov et al. 2012).

8.2. MPRAnator: a web-based tool for the design of massively parallel reporter assay experiments.

Here a set of web-based tools are introduced under the name MPRAnator (Georgakopoulos-Soares et al. 2017). Together these facilitate fast and precise design of MPRA experiments. These tools allow for the design of reporter assay sequences to investigate the regulatory roles of motifs and SNPs. MPRAnator is implemented in Python, Perl and Javascript and can be found in www.genomegeek.com and in www.sanger.ac.uk/science/tools/mpranator.

8.2.1. The MPRAnator Motif design tool.

The MPRAnator Motif Design tool is used to design sequences with motifs inserted alone or in combinations across the input sequences at intervals defined by the user. The program requires at least two inputs which are a set of FASTA scaffold sequences and a list of motifs in FASTA format, while it can also accept additional, optional inputs. The output of MPRAnator Motif design tool is a set of FASTA sequences, which have the motifs that the user selected integrated in the designed sequences (Figure 8.2).

A set of optional inputs is available to help the user design the sequences at a finer level of precision. The user can decide to include or exclude identifiable barcodes from the design, can define the minimum Levenshtein distance between any pair of barcodes and also restrict the range of the barcode GC content. The number of barcode tags per designed sequence can also be added, which results in the generation of multiple replicates per designed sequence. These options allow the user to optimise the experiment and limit the effect of other unwanted variables which could influence the results of the experiment.

Moreover, the user can adjust the frequency at which motifs are inserted into the input sequences and a set of restriction parameters control the distance relative to the start and the end of the input sequences, at which motifs can be placed at (Figure 8.2). If the user inputs more than one motif, output sequences will be generated for each motif independently as well as in combinations. The input sequences can also be reverse complemented to test for putative enhancer activity.

On the website, the input options described have been set to sensible default values. However, the user can decide which settings are best suited for their experiment. Vector integration is a crucial step during the design of MPRA experiments. To help with this step, the user can decide to include restriction sites, adaptor sequences or other sequences that would help their experiment. Furthermore, for each output sequence the header information includes all the necessary information regarding the motifs inserted and their corresponding positions, in the sequence the barcode information and other subcomponents of the scaffold that the user has decided to include as well as their order in the final constructed sequences. Importantly, if restriction site motifs used in the experiment are identified in any generated sequences, these are also reported in the header of the corresponding sequence and the user can decide to discard such sequences. The positions at which motifs are inserted are colour-marked to aid with visualisation.

Finally, a drag and drop interface allows modular design by placing the subcomponents of the construct such as adaptors, restriction sites, barcodes and designed oligonucleotides at the order that the user prefers (Figure 8.3). As a result of its flexibility, the MPRAnator Motif Design tool could be implemented in other protocols of reporter assays.

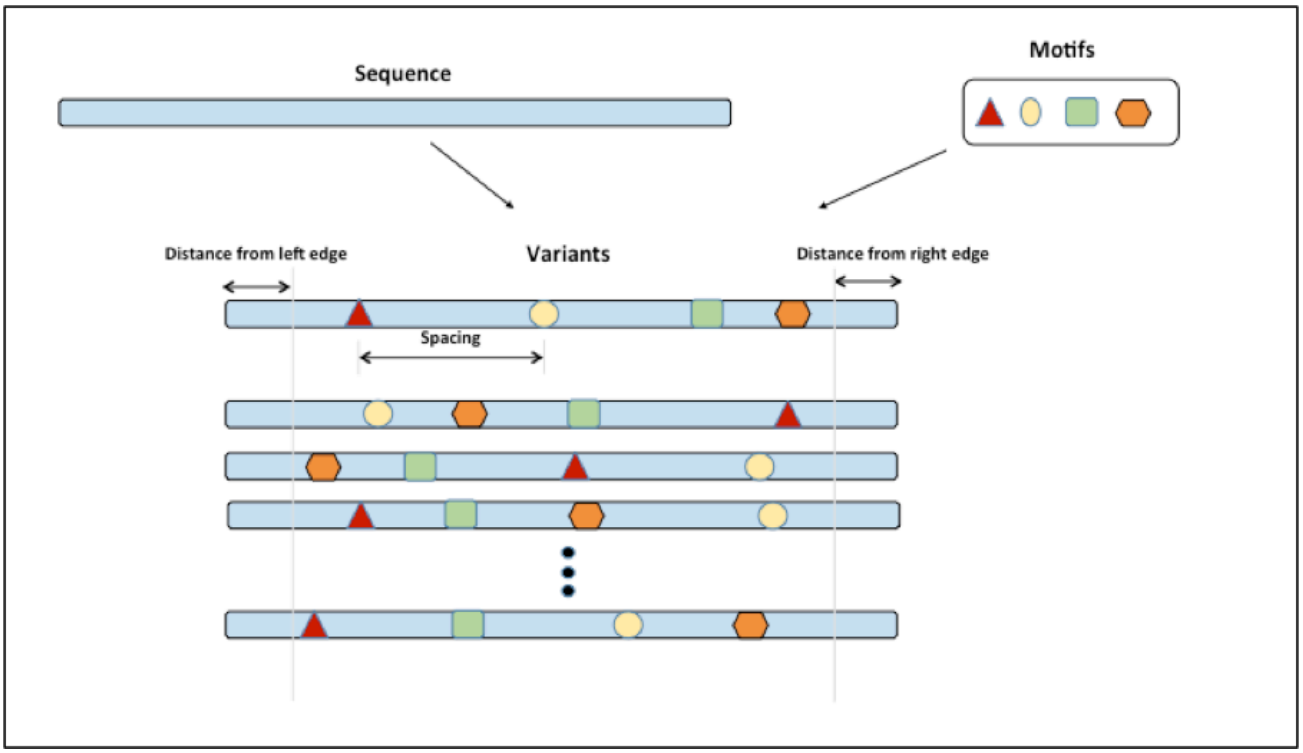


Figure 8.2: Schematic representation of the MPRAnator Motif design tool.

A maximum of four DNA motifs are integrated in the background sequences using all possible permutations. The distance from the edges and minimum and maximum spacing between motifs restrict the positioning of motifs in the sequences. The interval between motif positions determines the frequency of motif insertions in consecutive sequences (e.g. inserted every X nucleotides in the background sequence).

8.2.2. MPRAnator SNP design tool.

The MPRAnator SNP design tool facilitates the design of sequences to investigate the role of regulatory variants in gene expression independently or in combinations. Two inputs are necessary for this tool, a set of FASTA sequences that can be inserted or uploaded and an associated list of SNPs in the form of a variant call file (VCF). Regarding the VCF format, the tool supports up to 12 columns for each locus being inputted, but only uses the information present in the first 5 columns. For each sequence in the FASTA file, the associated SNPs will be identified and substituted. If multiple SNPs are present in a single inputted oligonucleotide sequence then all combinations are generated, which

allows the investigation of combinatorial effects. In addition, insertions and deletions in the VCF format are also permitted (Figure 8.3). Because insertions and deletions would result in the generation of output sequences with unequal lengths, in the case of insertions one end is trimmed to reduce the sequence length, whereas in the case of deletions adenines are added to the end of the generated sequences to increase the sequence length and to ensure that the final products all have the size length (Figure 8.3).

Similarly, to the MPRAnator Motif design tool, additional optional inputs are available to the user to provide extra parameters and ensure optimal experimental design. More specifically, the user can select to include in the construct uniquely identifiable barcodes that are incorporated in the output sequences. The barcode design includes multiple options; firstly, the barcode length can be decided upon (zero length, results in no barcode generation), the minimum Levenshtein distance between any pair of barcodes generated and the range of GC content which barcodes are allowed to have. Furthermore, the user can decide to substitute only one SNP at a time per sequence, or perform combinations as well, for sequences with multiple SNPs present.

The design of the vector is crucial in MPRA experiments. This tool allows the inclusion of restriction sites, adaptor sequences and other additional sequences that the user can decide to include in the design of his experiment. The header of each produced sequence contains information regarding the SNP name, position and sequence that was used and the order of the subcomponents in the designed construct. If the user has used restriction sites, a motif finding algorithm is implemented to report the output sequences which contain restriction sites and could affect the experiment. A drag and drop interface allows modular design by placing the subcomponents of the construct such as adaptors, restriction sites, barcodes, designed oligonucleotides or other inserted sequences at the order that the user prefers and is best suited to his experiment (Figure 8.4).

Importantly, the tool can accept a single scaffold of FASTA sequences without an associated VCF file with SNP information. This is useful when sequences with unknown

functionalities are needed to be investigated, with no interest in the regulatory role of SNPs.

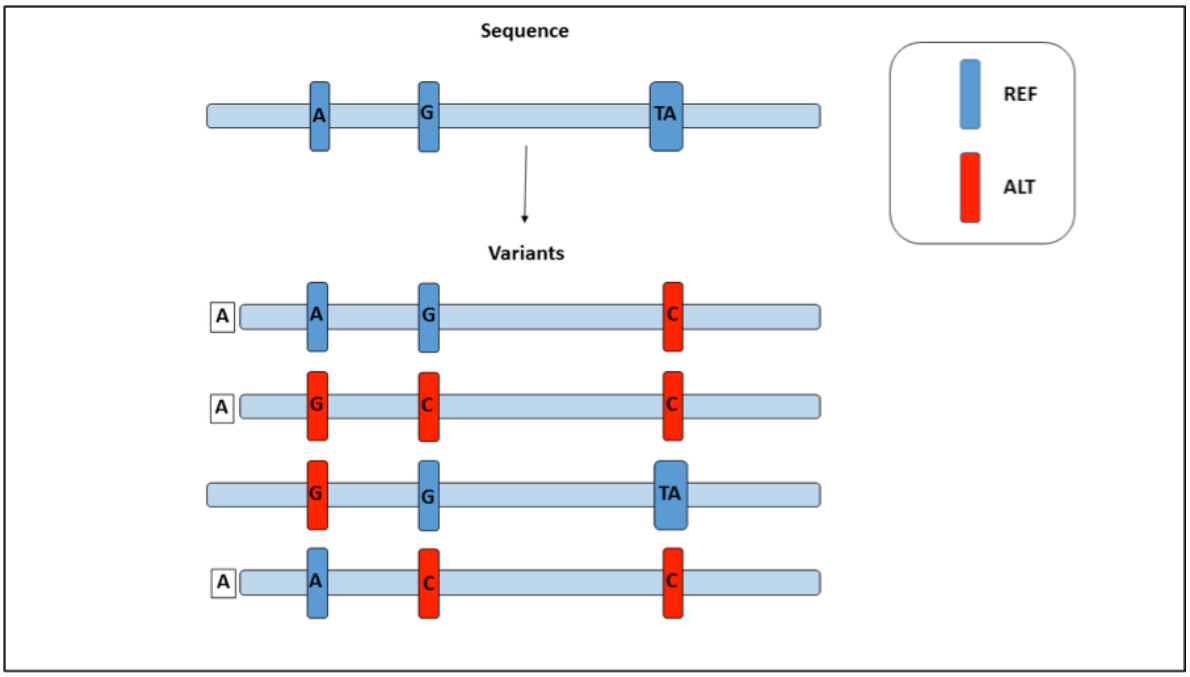


Figure 8.3: Schematic representation of the MPRAnator SNP design tool.

For each SNP position, a set of sequences with all available variants are designed. If multiple SNPs are present within the same sequence fragment, then the user can select to generate designed sequences with combinations of SNPs present or with single variant changes. For deletions, adenines are inserted at the end of the sequence and for insertions the sequence is trimmed to maintain the same length across output sequences.

8.2.3. MPRAnator Transmutation tool.

The MPRAnator Transmutation tool can be implemented to deconstruct information stored in a set of sequences or motifs that are inserted or uploaded in FASTA format. The tool has four options from which the user can pick. These are: i) scrambling the inputted sequences, therefore controlling for nucleotide composition, ii) reversing the sequence, which would destroy transcription factor binding sites, iii) complementing a set of

sequences, which would also control for nucleotide composition and would destroy transcription factor binding sites, iv) introducing a set of random mutations at each sequence in the set of inserted sequences. In particular, multiple options can be selected together, such as reverse and complementing a sequence, to study enhancer function or also adding random mutations. This tool can be used for the generation of negative controls that can be inputted in MPRAnator Motif design and MPRAnator SNP design tools.

The input format of this tool is a set of FASTA sequences and it will output the deconstructed sequences in FASTA format. The headers contain all the relevant information which include the number of random mutations generated for each sequence, or information regarding scrambling, reversing or complementing a sequence. Finally, mutated nucleotides for each output sequence have been colour-marked for visualisation purposes.

8.2.4. PWM Seq-Gen tool.

Regulatory sequence information is commonly stored in k-mer motifs or Positional Weight Matrices (PWMs), each of which has its own advantages and disadvantages. The function of the PWM Seq-Gen tool is to convert PWMs into k-mer sequence motifs. Because the tools described earlier use motifs in the k-mer format, any PWMs have to be converted before they can be incorporated into the MPRA design. For each PWM the format must include a header designated with “>” and four rows, each representing one nucleotide letter (A, C, G, T respectively). Columns in the matrix must be tab or space separated and they represent the position along the PWM motif.

Multiple optional variables are available to the user. Firstly, the user can select to return all possible k-mers that can be generated from a single PWM, using a probability threshold, which is also set by the user. This is particularly useful, if the user is interested

to compare similar motifs, that could have distinct functions. Secondly, the user can select to generate probabilistic realisations of PWMs, which will be converted into k-mers, with the relevant information available in their header. If duplicates are generated in the process, there is an additional option to filter them out.

The output sequences / motifs are generated in FASTA format. As with the previously described tools, the header contains the information regarding the k-mer motif that has been generated and the type of conversion that took place.

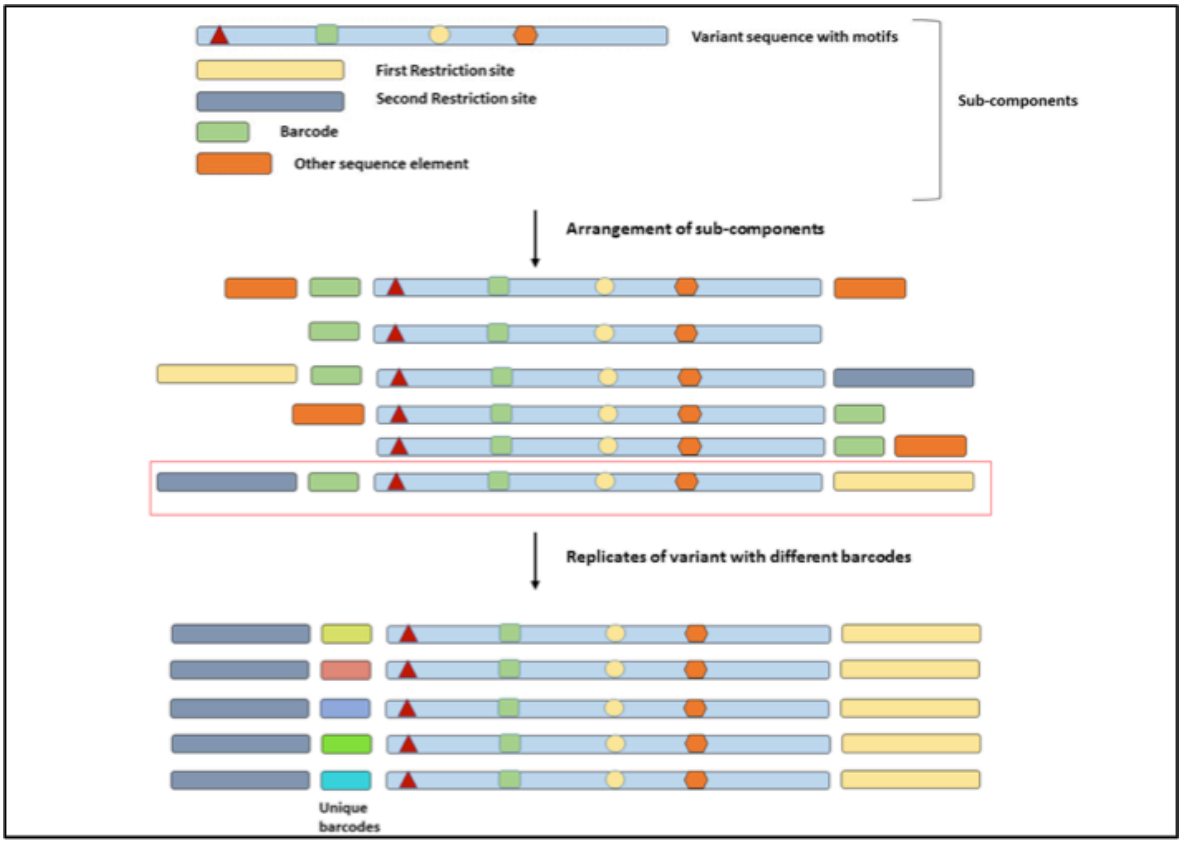


Figure 8.4: Modular design is implemented in MPRAnator for the final output.

Sub-components can be placed in any order therefore allowing for usage in different reporter assay experiments and increasing the flexibility of the design. The number of replicates for each sequence is decided upon by the user and each replicate has a distinct barcode tag.

8.3. Discussion.

Transcription factors control gene expression levels, by binding to TFBSs at promoter and enhancer regions (Zinzen et al. 2009). Cis-regulatory regions contain multiple TFBSs, at which transcription factors can bind in homotypic and heterotypic clusters to both activate and repress transcription of the associated genes. In addition, the vast majority of disease-causing variants are found in non-coding regions of the genome (Maurano et al. 2012) and their effects in gene regulation have been difficult to elucidate in the majority of cases.

MPRA technology allows for high-throughput investigations using reporter assays, and could provide insight into the regulatory code of the genome as well as the effects of disease-causing regulatory variants among many other usages. In addition, the generation of synthetic regulatory sequences, which have not been previously designed and are not present in nature, could further our understanding of transcription regulation. As beautifully captured by the renowned physicist Richard P. Feynman “What I cannot create, I do not understand.”.

MPRAnator is the first available tool that allows flexible and fast design of MPRA experiments with precision (Georgakopoulos-Soares et al. 2017). As the DNA sequencing and synthesis costs drop, MPRA technology could become widely used; however the availability of relevant tools to facilitate the design of these experiments is key to its success and implementation.

CHAPTER NINE

9. Bibliography.

Adachi M, Tsujimoto Y. 1990. Potential Z-DNA elements surround the breakpoints of chromosome translocation within the 5' flanking region of bcl-2 gene. *Oncogene* **5**: 1653–1657.

Agarwal T, Lalwani MK, Kumar S, Roy S, Chakraborty TK, Sivasubbu S, Maiti S. 2014. Morphological effects of G-quadruplex stabilization using a small molecule in zebrafish. *Biochemistry* **53**: 1117–1124.

Agazie YM, Burkholder GD, Lee JS. 1996. Triplex DNA in the nucleus: direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. *Biochem J* **316 (Pt 2)**: 461–466.

Agazie YM, Lee JS, Burkholder GD. 1994. Characterization of a new monoclonal antibody to triplex DNA and immunofluorescent staining of mammalian chromosomes. *J Biol Chem* **269**: 7019–7023.

Agrawal P, Lin C, Mathad RI, Carver M, Yang D. 2014. The Major G-Quadruplex Formed in the Human BCL-2 Proximal Promoter Adopts a Parallel Structure with a 13-nt Loop in K Solution. *J Am Chem Soc* **136**: 1750–1753.

Akgün E, Zahn J, Baumes S, Brown G, Liang F, Romanienko PJ, Lewis S, Jasin M. 1997. Palindrome resolution and recombination in the mammalian germ line. *Mol Cell Biol* **17**: 5559–5570.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**: 415–421.

Alexandrov L, Kim J, Haradhvala NJ, Huang MN, Ng AWT, Boot A, Covington KR, Gordenin DA, Bergstrom E, Lopez-Bigas N, et al. 2018. The Repertoire of Mutational Signatures in Human Cancer. http://dx.doi.org/10.1101/322859.

Amrane S, Adrian M, Heddi B, Serero A, Nicolas A, Mergny J-L, Phan AT. 2012. Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite. *J Am Chem Soc* **134**: 5807–5816.

Andrianova MA, Bazykin GA, Nikolaev SI, Seplyarskiy VB. 2017. Human mismatch repair system balances mutation rates between strands by removing more mismatches from the lagging strand. *Genome Res* **27**: 1336–1343.

Armas P, David A, Calcaterra NB. 2017. Transcriptional control by G-quadruplexes: In vivo roles and perspectives for specific intervention. *Transcription* **8**: 21–25.

Axford MM, Wang Y-H, Nakamori M, Zannis-Hadjopoulos M, Thornton CA, Pearson CE. 2013. Detection of slipped-DNAs at the trinucleotide repeats of the myotonic dystrophy type I disease locus in patient tissues. *PLoS Genet* **9**: e1003866.

Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, Stefanov S, Olsh A, Jakupciak JP, Dean M, et al. 2006. Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res* **34**: 2663–2675.

Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, Abeysinghe SS, O'Connell CD, Cooper DN, Wells RD. 2004. Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc Natl Acad Sci U S A* **101**: 14162–14167.

Bacolla A, Tainer JA, Vasquez KM, Cooper DN. 2016. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res* **44**: 5673–5688.

Bacolla A, Wang G, Jain A, Chuzhanova NA, Cer RZ, Collins JR, Cooper DN, Bohr VA, Vasquez KM. 2011. Non-B DNA-forming Sequences and WRN Deficiency Independently Increase the Frequency of Base Substitution in Human Cells. *J Biol Chem* **286**: 10017–10026.

Bacolla A, Wells RD. 2004. Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem* **279**: 47411–47414.

Bagshaw ATM. 2017. Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. *Genome Biol Evol* **9**: 2428–2443.

Baraniak AP, Lasda EL, Wagner EJ, Garcia-Blanco MA. 2003. A stem structure in fibroblast growth factor receptor 2 transcripts mediates cell-type-specific splicing by approximating intronic control elements. *Mol Cell Biol* **23**: 9327–9337.

Bartas M, Brázda V, Karlický V, Červeň J, Pečinka P. 2018. Bioinformatics analyses and in vitro evidence for five and six stacked G-quadruplex forming sequences. *Biochimie* **150**: 70–75.

Baskett W, Spencer M, Shyu C-R. 2017. Efficient GPU-accelerated extraction of imperfect inverted repeats from DNA sequences. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* http://dx.doi.org/10.1109/bibm.2017.8217700.

Bay DH, Busch A, Lisdat F, Iida K, Ikebukuro K, Nagasawa K, Karube I, Yoshida W. 2017. Identification of G-quadruplex structures that possess transcriptional regulating functions in the Dele and Cdc6 CpG islands. *BMC Mol Biol* **18**: 17.

Beaudoin J-D, Jodoin R, Perreault J-P. 2013. In-line probing of RNA G-quadruplexes. *Methods* **64**: 79–87.

Beaudoin J-D, Perreault J-P. 2010. 5'-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res* **38**: 7022–7036.

Beier S, Thiel T, Münch T, Scholz U, Mascher M. 2017. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**: 2583–2585.

Belotserkovskii BP, De Silva E, Tornaletti S, Wang G, Vasquez KM, Hanawalt PC. 2007. A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J Biol Chem* **282**: 32433–32441.

Benabou S, Ferreira R, Aviñó A, González C, Lyonnais S, Solà M, Eritja R, Jaumot J, Gargallo R. 2014. Solution equilibria of cytosine- and guanine-rich sequences near the

promoter region of the n-myc gene that contain stable hairpins within lateral loops. *Biochim Biophys Acta* **1840**: 41–52.

Benham CJ, Savitt AG, Bauer WR. 2002. Extrusion of an imperfect palindrome to a cruciform in superhelical DNA: complete determination of energetics using a statistical mechanical model. *J Mol Biol* **316**: 563–581.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, Vitale L, Pelleri MC, Tassani S, Piva F, et al. 2013. An estimation of the number of cells in the human body. *Ann Hum Biol* **40**: 463–471.

Bidichandani SI, Ashizawa T, Patel PI. 1998. The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. *Am J Hum Genet* **62**: 111–121.

Biffi G, Tannahill D, McCafferty J, Balasubramanian S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat Chem* **5**: 182–186.

Biffi G, Tannahill D, Miller J, Howat WJ, Balasubramanian S. 2014. Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. *PLoS One* **9**: e102711.

Bochman ML, Paeschke K, Zakian VA. 2012. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**: 770–780.

Boiteux S, Costa de Oliveira R, Laval J. 1985. The Escherichia coli O6-methylguanine-DNA methyltransferase does not repair promutagenic O6-methylguanine residues when present in Z-DNA. *J Biol Chem* **260**: 8711–8715.

Borel C, Migliavacca E, Letourneau A, Gagnebin M, Béna F, Sailani MR, Dermitzakis ET, Sharp AJ, Antonarakis SE. 2012. Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. *Hum Mutat* **33**: 1302–1309.

Boyer A-S, Grgurevic S, Cazaux C, Hoffmann J-S. 2013. The human specialized DNA polymerases and non-B DNA: vital relationships to preserve genome integrity. *J Mol Biol* **425**: 4767–4781.

Branzei D, Foiani M. 2010. Leaping forks at inverted repeats. *Genes Dev* **24**: 5–9.

Bratic A, Larsson N-G. 2013. The role of mitochondria in aging. *J Clin Invest* **123**: 951–957.

Brázda V, Kolomazník J, Lýsek J, Hároníková L, Coufal J, Šťastný J. 2016. Palindrome analyser - A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem Biophys Res Commun* **478**: 1739–1745.

Brooks TA, Hurley LH. 2010. Targeting MYC Expression through G-Quadruplexes. *Genes Cancer* **1**: 641–649.

Brosh RM Jr, Majumdar A, Desai S, Hickson ID, Bohr VA, Seidman MM. 2001. Unwinding of a DNA triple helix by the Werner and Bloom syndrome helicases. *J Biol Chem* **276**: 3024–3030.

Bugaut A, Balasubramanian S. 2008. A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry* **47**: 689–697.

Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. 2006. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* **34**: 5402–5415.

Buske FA, Bauer DC, Mattick JS, Bailey TL. 2012. Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res* **22**: 1372–1381.

Buske FA, Bauer DC, Mattick JS, Bailey TL. 2013. Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. *Bioinformatics* **29**: 1895–1897.

Campbell NH, Parkinson GN. 2007. Crystallographic studies of quadruplex nucleic acids. *Methods* **43**: 252–263.

Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD, - ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net. 2017. Pan-cancer analysis of whole genomes. http://dx.doi.org/10.1101/162784.

Campuzano V, Montermini L, Moltò MD, Pianese L, Cossée M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A, et al. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**: 1423–1427.

Carè A, Cianetti L, Giampaolo A, Sposi NM, Zappavigna V, Mavilio F, Alimena G, Amadori S, Mandelli F, Peschle C. 1986. Translocation of c-myc into the immunoglobulin heavy-chain locus in human acute B-cell leukemia. A molecular analysis. *EMBO J* **5**: 905–911.

Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, Bacolla A, Collins JR, Stephens RM. 2011. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* **39**: D383–91.

Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT, et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41**: D94–D100.

Chaires JB, Trent JO, Gray RD, Dean WL, Buscaglia R, Thomas SD, Miller DM. 2014. An improved model for the hTERT promoter quadruplex. *PLoS One* **9**: e115580.

Chambers VS, Marsico G, Boutell JM, Di Antonio M, Smith GP, Balasubramanian S. 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**: 877–881.

Champ PC, Maurice S, Vargason JM, Camp T, Ho PS. 2004. Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation. *Nucleic Acids Res* **32**: 6501–6510.

Chang HHY, Pannunzio NR, Adachi N, Lieber MR. 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol* **18**: 495–506.

Chen L, Chen J-Y, Zhang X, Gu Y, Xiao R, Shao C, Tang P, Qian H, Luo D, Li H, et al. 2017. R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol Cell* **68**: 745–757.e5.

Chen MC, Tippana R, Demeshkina NA, Murat P, Balasubramanian S, Myong S, Ferré-D'Amaré AR. 2018. Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature* **558**: 465–469.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741–754.

Chen RH, Maher VM, Brouwer J, van de Putte P, McCormick JJ. 1992. Preferential repair and strand-specific repair of benzo[a]pyrene diol epoxide adducts in the HPRT gene of diploid human fibroblasts. *Proc Natl Acad Sci U S A* **89**: 5413–5417.

Chen Y, Yang D. 2012. Sequence, stability, and structure of G-quadruplexes and their interactions with drugs. *Curr Protoc Nucleic Acid Chem* **Chapter 17**: Unit17.5.

Chetsanga CJ, Boyd V, Peterson L, Rushlow K. 1975. Single-stranded regions in DNA of old mice. *Nature* **253**: 130–131.

Che T, Wang Y-Q, Huang Z-L, Tan J-H, Huang Z-S, Chen S-B. 2018. Natural Alkaloids and Heterocycles as G-Quadruplex Ligands and Potential Anticancer Agents. *Molecules* **23**. http://dx.doi.org/10.3390/molecules23020493.

Chew DSH, Choi KP, Leung M-Y. 2005. Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. *Nucleic Acids Res* **33**: e134.

Christensen LA, Finch RA, Booker AJ, Vasquez KM. 2006. Targeting oncogenes to improve breast cancer chemotherapy. *Cancer Res* **66**: 4089–4094.

Cogoi S, Rapozzi V, Cauci S, Xodo LE. 2017. Critical role of hnRNP A1 in activating KRAS transcription in pancreatic cancer cells: A molecular mechanism involving G4 DNA. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1861**: 1389–1398.

Collins J. 1981. Instability of palindromic DNA in Escherichia coli. *Cold Spring Harb Symp Quant Biol* **45 Pt 1**: 409–416.

Collins J, Volckaert G, Nevers P. 1982. Precise and nearly-precise excision of the symmetrical inverted repeats of Tn5; common features of recA-independent deletion events in Escherichia coli. *Gene* **19**: 139–146.

Cooney M, Czernuszewicz G, Postel EH, Flint SJ, Hogan ME. 1988. Site-specific oligonucleotide binding represses transcription of the human c-myc gene in vitro. *Science* **241**: 456–459.

Cooper DN, Krawczak M. 1991. Mechanisms of insertional mutagenesis in human genes causing genetic disease. *Hum Genet* **87**: 409–415.

Costello E, Sahli R, Hirt B, Beard P. 1995. The mismatched nucleotides in the 5'-terminal hairpin of minute virus of mice are required for efficient viral DNA replication. *J Virol* **69**: 7489–7496.

Cox R, Mirkin SM. 1997. Characteristic enrichment of DNA repeats in different genomes. *Proc Natl Acad Sci U S A* **94**: 5237–5242.

d'Adda di Fagagna F, di Fagagna FD. 2008. Molecular mechanisms of cellular senescence. *Eur J Cancer Suppl* **6**: 35.

Dai J, Chen D, Jones RA, Hurley LH, Yang D. 2006. NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Res* **34**: 5133–5144.

Damas J, Carneiro J, Gonçalves J, Stewart JB, Samuels DC, Amorim A, Pereira F. 2012. Mitochondrial DNA deletions are associated with non-B DNA conformations. *Nucleic Acids Res* **40**: 7606–7621.

Darlow JM, Leach DR. 1998. Secondary structures in d(CGG) and d(CCG) repeat tracts. *J Mol Biol* **275**: 3–16.

Davies H, Glodzik D, Morganella S, Yates LR, Staaf J, Zou X, Ramakrishna M, Martin S, Boyault S, Sieuwerts AM, et al. 2017. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* **23**: 517–525.

Davis JT. 2004. G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. *Angew Chem Int Ed Engl* **43**: 668–698.

Davis TL, Firulli AB, Kinniburgh AJ. 1989. Ribonucleoprotein and protein factors bind to an H-DNA-forming c-myc DNA element: possible regulators of the c-myc gene. *Proceedings of the National Academy of Sciences* **86**: 9682–9686.

Dayn A, Malkhosyan S, Mirkin SM. 1992. Transcriptionally driven cruciform formationin vivo. *Nucleic Acids Res* **20**: 5991–5997.

de Boer JG, Ripley LS. 1984. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences. *Proc Natl Acad Sci U S A* **81**: 5528–5531.

DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, et al. 2011. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**: 245–256.

Del Mundo IMA, Zewail-Foote M, Kerwin SM, Vasquez KM. 2017. Alternative DNA structure formation in the mutagenic human c-MYC promoter. *Nucleic Acids Res* **45**: 4929–4943.

Denissenko MF, Pao A, Pfeifer GP, Tang M. 1998. Slow repair of bulky DNA adducts along the nontranscribed strand of the human p53 gene may explain the strand bias of transversion mutations in cancers. *Oncogene* **16**: 1241–1247.

De S, Michor F. 2011. DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol* **18**: 950–955.

Di Antonio M, Biffi G, Mariani A, Raiber E-A, Rodriguez R, Balasubramanian S. 2012. Selective RNA versus DNA G-quadruplex targeting by in situ click chemistry. *Angew Chem Int Ed Engl* **51**: 11073–11078.

Ditlevson JV, Tornaletti S, Belotserkovskii BP, Teijeiro V, Wang G, Vasquez KM, Hanawalt PC. 2008. Inhibitory effect of a short Z-DNA forming sequence on transcription elongation by T7 RNA polymerase. *Nucleic Acids Res* **36**: 3163–3170.

Doktycz MJ, Benight AS, Sheardy RD. 1990. Energetics of B-Z junction formation in a sixteen base-pair duplex DNA. *J Mol Biol* **212**: 3–6.

Dong Y, Yang Z, Liu D. 2014. DNA nanotechnology based on i-motif structures. *Acc Chem Res* **47**: 1853–1860.

Dubrova YE, Jeffreys AJ, Malashenko AM. 1993. Mouse minisatellite mutations induced by ionizing radiation. *Trends Genet* **9**: 379.

Dubrova YE, Plumb M, Brown J, Jeffreys AJ. 1998. Radiation-induced germline instability at minisatellite loci. *Int J Radiat Biol* **74**: 689–696.

Dumelie JG, Jaffrey SR. 2017. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. *Elife* **6**. http://dx.doi.org/10.7554/eLife.28306.

Du X, Gertz EM, Wojtowicz D, Zhabinskaya D, Levens D, Benham CJ, Schäffer AA, Przytycka TM. 2014. Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic Acids Res* **42**: 12367–12379.

Dzatko S, Krafcikova M, Hänsel-Hertsch R, Fessl T, Fiala R, Loja T, Krafcik D, Mergny J-L, Foldynova-Trantirkova S, Trantirek L. 2018. Evaluation of the Stability of DNA i-Motifs in the Nuclei of Living Mammalian Cells. *Angew Chem Int Ed Engl* **57**: 2165–2169.

Echlin-Bell DR, Smith LL, Li L, Strissel PL, Strick R, Gupta V, Banerjee J, Larson R, Relling MV, Raimondi SC, et al. 2003. Polymorphisms in the MLL breakpoint cluster region (BCR). *Hum Genet* **113**: 80–91.

Eddy J, Vallur AC, Varma S, Liu H, Reinhold WC, Pommier Y, Maizels N. 2011. G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Res* **39**: 4975–4983.

Ellison MJ, Kelleher RJ, Wang AH, Habener JF, Rich A. 1985. Sequence-dependent energetics of the B-Z transition in supercoiled DNA containing nonalternating purine-pyrimidine sequences. *Proceedings of the National Academy of Sciences* **82**: 8320–8324.

Eykelenboom JK, Blackwood JK, Okely E, Leach DRF. 2008. SbcCD causes a double-strand break at a DNA palindrome in the Escherichia coli chromosome. *Mol Cell* **29**: 644–651.

Faruqi AF, Datta HJ, Carroll D, Seidman MM, Glazer PM. 2000. Triple-Helix Formation Induces Recombination in Mammalian Cells via a Nucleotide Excision Repair-Dependent Pathway. *Mol Cell Biol* **20**: 990–1000.

Feigon J, Wang AH, van der Marel GA, van Boom JH, Rich A. 1985. Z-DNA forms without an alternating purine-pyrimidine sequence in solution. *Science* **230**: 82–84.

Felsenfeld G, Davies DR, Rich A. 1957. Formation of a Three-stranded polynucleotide molecule. *J Am Chem Soc* **79**: 2023–2024.

Frank-Kamenetskii MD, Mirkin SM. 1995. Triplex DNA structures. *Annu Rev Biochem* **64**: 65–95.

Fredriksson NJ, Ny L, Nilsson JA, Larsson E. 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**: 1258–1263.

Freund AM, Bichara M, Fuchs RP. 1989. Z-DNA-forming sequences are spontaneous deletion hot spots. *Proc Natl Acad Sci U S A* **86**: 7465–7469.

Fry M, Loeb LA. 1999. Human werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)n. *J Biol Chem* **274**: 12797–12802.

Gadgil R, Barthelemy J, Lewis T, Leffak M. 2017. Replication stalling and DNA microsatellite instability. *Biophys Chem* **225**: 38–48.

Garcia-Diaz M, Kunkel TA. 2006. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem Sci* **31**: 206–214.

Garner MM, Felsenfeld G. 1987. Effect of Z-DNA on nucleosome placement. *J Mol Biol* **196**: 581–590.

Gatalica Z, Vranic S, Xiu J, Swensen J, Reddy S. 2016. High microsatellite instability (MSI-H) colorectal carcinoma: a brief review of predictive biomarkers in the era of personalized medicine. *Fam Cancer* **15**: 405–412.

Gebhardt F, Zänker KS, Brandt B. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* **274**: 13176–13180.

Gehring K, Leroy JL, Guéron M. 1993. A tetrameric DNA structure with protonated cytosine.cytosine base pairs. *Nature* **363**: 561–565.

Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**: 445–477.

Georgakopoulos-Soares I, Jain N, Gray JM, Hemberg M. 2017. MPRAnator: a web-based tool for the design of massively parallel reporter assay experiments. *Bioinformatics* **33**: 137–138.

Georgakopoulos-Soares I, Morganella S, Jain N, Hemberg M, Nik-Zainal S. 2018. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res*. http://dx.doi.org/10.1101/gr.231688.117.

Gessner RV, Frederick CA, Quigley GJ, Rich A, Wang AH. 1989. The molecular structure of the left-handed Z-DNA double helix at 1.0 angstrom atomic resolution. Geometry, conformation, and ionic interactions of d(CGCGCG). http://dx.doi.org/10.2210/pdb1dcg/pdb.

Ghosh A, Bansal M. 2003. A glossary of DNA structures from A to Z. *Acta Crystallogr D Biol Crystallogr* **59**: 620–626.

Ginno PA, Lim YW, Lott PL, Korf I, Chédin F. 2013. GC skew at the 5’ and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* **23**: 1590–1600.

Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell* **45**: 814–825.

Gomez D, Lemarteleur T, Lacroix L, Mailliet P, Mergny J-L, Riou J-F. 2004. Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. *Nucleic Acids Res* **32**: 371–379.

Goñi JR, de la Cruz X, Orozco M. 2004. Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res* **32**: 354–360.

Goñi JR, Vaquerizas JM, Dopazo J, Orozco M. 2006. Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. *BMC Genomics* **7**: 63.

Gordenin DA, Lobachev KS, Degtyareva NP, Malkova AL, Perkins E, Resnick MA. 1993. Inverted DNA repeats: a source of eukaryotic genomic instability. *Mol Cell Biol* **13**: 5315–5322.

Gotter AL, Shaikh TH, Budarf ML, Rhodes CH, Emanuel BS. 2004. A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2. *Hum Mol Genet* **13**: 103–115.

Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. 2006. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**: 2187–2198.

Grote P, Herrmann BG. 2013. The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA Biol* **10**: 1579–1585.

Guédin A, Gros J, Alberti P, Mergny J-L. 2010. How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res* **38**: 7858–7868.

Guo M, Hundseth K, Ding H, Vidhyasagar V, Inoue A, Nguyen C-H, Zain R, Lee JS, Wu Y. 2015. A distinct triplex DNA unwinding activity of ChlR1 helicase. *J Biol Chem* **290**: 5174–5189.

Gurung SP, Schwarz C, Hall JP, Cardin CJ, Brazier JA. 2015. The importance of loop length on the stability of i-motif structures. *Chem Commun* **51**: 5630–5632.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, et al. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* **48**: 22–29.

Hagerman PJ, Hagerman RJ. 2004. The fragile-X premutation: a maturing perspective. *Am J Hum Genet* **74**: 805–816.

Halder K, Wieland M, Hartig JS. 2009. Predictable suppression of gene expression by 5′-UTR-based RNA quadruplexes. *Nucleic Acids Res* **37**: 6811–6817.

Hall K, Cruz P, Tinoco I Jr, Jovin TM, van de Sande JH. 1984. “Z-RNA”--a left-handed RNA double helix. *Nature* **311**: 584–586.

Haluska FG, Tsujimoto Y, Croce CM. 1988. The t(8;14) breakpoint of the EW 36 undifferentiated lymphoma cell line lies 5’ of MYC in a region prone to involvement in endemic Burkitt's lymphomas. *Nucleic Acids Res* **16**: 2077–2085.

Hamperl S, Bocek MJ, Saldivar JC, Swigut T, Cimprich KA. 2017. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170**: 774–786.e19.

Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat Rev Mol Cell Biol* **9**: 958–970.

Haniford DB, Pulleyblank DE. 1983a. Facile transition of poly[d(TG) x d(CA)] into a left-handed helix in physiological conditions. *Nature* **302**: 632–634.

Haniford DB, Pulleyblank DE. 1983b. The in-vivo occurrence of Z DNA. *J Biomol Struct Dyn* **1**: 593–609.

Hannan AJ. 2010. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for “missing heritability.” *Trends Genet* **26**: 59–65.

Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19**: 286–298.

Hänsel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, et al. 2016. G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**: 1267–1272.

Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. 2017. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* **18**: 279–284.

Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyannopoulos JA. 2010. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**: 139–144.

Hanzelmann S, Kuo C-C, Kalwa M, Wagner W, Costa IG. 2015. Triplex Domain Finder: Detection of Triple Helix Binding Domains in Long Non-Coding RNAs. http://dx.doi.org/10.1101/020297.

Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, et al. 2016. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**: 538–549.

Ha SC, Lowenhaupt K, Rich A, Kim Y-G, Kim KK. 2005. Crystal structure of a junction between B-DNA and Z-DNA reveals two extruded bases. *Nature* **437**: 1183–1186.

Hazel P, Huppert J, Balasubramanian S, Neidle S. 2004. Loop-length-dependent folding of G-quadruplexes. *J Am Chem Soc* **126**: 16405–16415.

Henderson A, Wu Y, Huang YC, Chavez EA, Platt J, Johnson FB, Brosh RM Jr, Sen D, Lansdorp PM. 2017. Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res* **45**: 6252.

Herbert AG, Rich A. 1993. A method to identify and characterize Z-DNA binding proteins using a linear oligodeoxynucleotide. *Nucleic Acids Res* **21**: 2669–2672.

Herbert A, Lowenhaupt K, Spitzner J, Rich A. 1995. Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA. *Proc Natl Acad Sci U S A* **92**: 7550–7554.

Herbert A, Rich A. 1996. The biology of left-handed Z-DNA. *J Biol Chem* **271**: 11595–11598.

Herbert A, Schade M, Lowenhaupt K, Alfken J, Schwartz T, Shlyakhtenko LS, Lyubchenko YL, Rich A. 1998. The Zalpha domain from human ADAR1 binds to the Z-DNA conformer of many different sequences. *Nucleic Acids Res* **26**: 3486–3493.

Hershman SG, Chen Q, Lee JY, Kozak ML, Yue P, Wang L-S, Johnson FB. 2008. Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res* **36**: 144–156.

Hewett PW, Daft EL, Laughton CA, Ahmad S, Ahmed A, Murray JC. 2006. Selective inhibition of the human tie-1 promoter with triplex-forming oligonucleotides targeted to Ets binding sites. *Mol Med* **12**: 8–16.

Hewish M, Lord CJ, Martin SA, Cunningham D, Ashworth A. 2010. Mismatch repair deficient colorectal cancer in the era of personalized treatment. *Nat Rev Clin Oncol* **7**: 197–208.

Hoang ML, Chen C-H, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, Douville C, Karchin R, et al. 2013. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med* **5**: 197ra102.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473–476.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**: 827–841.

Htun H, Dahlberg JE. 1988. Single strands, triple strands, and kinks in H-DNA. *Science* **241**: 1791–1796.

Huang D-S, Wang Z, He X-J, Diplas BH, Yang R, Killela PJ, Meng Q, Ye Z-Y, Wang W, Jiang X-T, et al. 2015. Recurrent TERT promoter mutations identified in a large-scale study of multiple tumour types are associated with increased TERT expression and telomerase activation. *Eur J Cancer* **51**: 969–976.

Huppert JL. 2010. Structure, location and interactions of G-quadruplexes. *FEBS J* **277**: 3452–3458.

Huppert JL, Balasubramanian S. 2007. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res* **35**: 406–413.

Huppert JL, Balasubramanian S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**: 2908–2916.

Imielinski M, Guo G, Meyerson M. 2017. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell* **168**: 460–472.e14.

Inagaki H, Ohye T, Kogo H, Kato T, Bolor H, Taniguchi M, Shaikh TH, Emanuel BS, Kurahashi H. 2009. Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. *Genome Res* **19**: 191–198.

Inoue F, Ahituv N. 2015. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**: 159–164.

Jain A, Bacolla A, Del Mundo IM, Zhao J, Wang G, Vasquez KM. 2013. DHX9 helicase is involved in preventing genomic instability induced by alternatively structured DNA in human cells. *Nucleic Acids Res* **41**: 10345–10357.

Jain A, Wang G, Vasquez KM. 2008. DNA triple helices: biological consequences and therapeutic potential. *Biochimie* **90**: 1117–1130.

Jalali S, Singh A, Maiti S, Scaria V. 2017. Genome-wide computational analysis of potential long noncoding RNA mediated DNA:DNA:RNA triplexes in the human genome. *J Transl Med* **15**: 186.

James PL, Brown T, Fox KR. 2003. Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+*GC and T*AT triplets. *Nucleic Acids Res* **31**: 5598–5606.

Jasin M, Rothstein R. 2013. Repair of Strand Breaks by Homologous Recombination. *Cold Spring Harb Perspect Biol* **5**: a012740–a012740.

Jaworski A, Blaho JA, Larson JE, Shimizu M, Wells RD. 1989. Tetracycline promoter mutations decrease non-B DNA structural transitions, negative linking differences and deletions in recombinant plasmids in Escherichia coli. *J Mol Biol* **207**: 513–526.

Jaworski A, Hsieh WT, Blaho JA, Larson JE, Wells RD. 1987. Left-handed DNA in vivo. *Science* **238**: 773–777.

Jenjaroenpun P, Kuznetsov VA. 2009. TTS mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics* **10 Suppl 3**: S9.

Jodoin R, Bauer L, Garant J-M, Mahdi Laaref A, Phaneuf F, Perreault J-P. 2014. The folding of 5'-UTR human G-quadruplexes possessing a long central loop. *RNA* **20**: 1129–1141.

Joos S, Falk MH, Lichter P, Haluska FG, Henglein B, Lenoir GM, Bornkamm GW. 1992. Variable breakpoints in Burkitt lymphoma cells with chromosomal t(8; 14) translocation separate c-myc and the IgH locus up to several hundred kb. *Hum Mol Genet* **1**: 625–632.

Kalwa M, Hänzelmann S, Otto S, Kuo C-C, Franzen J, Joussen S, Fernandez-Rebollo E, Rath B, Koch C, Hofmann A, et al. 2016. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res* **44**: 10631–10643.

Kamat MA, Bacolla A, Cooper DN, Chuzhanova N. 2016. A Role for Non-B DNA Forming Sequences in Mediating Microlesions Causing Human Inherited Disease. *Hum Mutat* **37**: 65–73.

Kato M, Hokabe S, Itakura S, Minoshima S, Lyubchenko YL, Gurkov TD, Okawara H, Nagayama K, Shimizu N. 2003. Interarm Interaction of DNA Cruciform Forming at a Short Inverted Repeat Sequence. *Biophys J* **85**: 402–408.

Kennedy GC, German MS, Rutter WJ. 1995. The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. *Nat Genet* **9**: 293–298.

Khuu P, Sandor M, DeYoung J, Ho PS. 2007. Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proc Natl Acad Sci U S A* **104**: 16528–16533.

Kikin O, D'Antonio L, Bagga PS. 2006. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**: W676–82.

Kim M, Kreig A, Lee C-Y, Rube HT, Calvert J, Song JS, Myong S. 2016. Quantitative analysis and prediction of G-quadruplex forming sequences in double-stranded DNA. *Nucleic Acids Res* **44**: 4807–4817.

Kim N, Jinks-Robertson S. 2012. Transcription as a source of genome instability. *Nat Rev Genet* **13**: 204–214.

Kim T-M, Laird PW, Park PJ. 2013. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* **155**: 858–868.

Kim YG, Lowenhaupt K, Schwartz T, Rich A. 1999. The interaction between Z-DNA and the Zab domain of double-stranded RNA adenosine deaminase characterized using fusion nucleases. *J Biol Chem* **274**: 19081–19086.

Kolpakov R, Bana G, Kucherov G. 2003. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res* **31**: 3672–3678.

Koshlap KM, Schultze P, Brunar H, Dervan PB, Feigon J. 1997. Solution structure of an intramolecular DNA triplex containing an N7-glycosylated guanine which mimics a protonated cytosine. *Biochemistry* **36**: 2659–2668.

Kouzine F, Wojtowicz D, Yamane A, Resch W, Kieffer-Kwon K-R, Bandle R, Nelson S, Nakahashi H, Awasthi P, Feigenbaum L, et al. 2013. Global regulation of promoter melting in naive lymphocytes. *Cell* **153**: 988–999.

Kovtun IV, Liu Y, Bjoras M, Klungland A, Wilson SH, McMurray CT. 2007. OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. *Nature* **447**: 447–452.

Krasilnikova MM, Kireeva ML, Petrovic V, Knijnikova N, Kashlev M, Mirkin SM. 2007. Effects of Friedreich's ataxia (GAA)nmiddle dot(TTC)n repeats on RNA synthesis and stability. *Nucleic Acids Res* **35**: 1075–1084.

Krasilnikova MM, Samadashwily GM, Krasilnikov AS, Mirkin SM. 1998. Transcription through a simple DNA repeat blocks replication elongation. *EMBO J* **17**: 5095–5102.

Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM. 1999. Large-scale effects of transcriptional DNA supercoiling in vivo. *J Mol Biol* **292**: 1149–1160.

Kubota M, Tran C, Spitale RC. 2015. Progress and challenges for chemical probing of RNA structure inside living cells. *Nat Chem Biol* **11**: 933–941.

Kurahashi H. 2001. Long AT-rich palindromes and the constitutional t(11;22) breakpoint. *Hum Mol Genet* **10**: 2605–2617.

Kurahashi H, Inagaki H, Kato T, Hosoba E, Kogo H, Ohye T, Tsutsumi M, Bolor H, Tong M, Emanuel BS. 2009. Impaired DNA replication prompts deletions within palindromic sequences, but does not induce translocations in human cells. *Hum Mol Genet* **18**: 3397–3406.

Kurahashi H, Inagaki H, Ohye T, Kogo H, Kato T, Emanuel BS. 2006. Palindrome-mediated chromosomal translocations in humans. *DNA Repair* **5**: 1136–1145.

Kurahashi H, Inagaki H, Yamada K, Ohye T, Taniguchi M, Emanuel BS, Toda T. 2004. Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. *J Biol Chem* **279**: 35377–35383.

Kwok CK, Merrick CJ. 2017. G-Quadruplexes: Prediction, Characterization, and Biological Application. *Trends Biotechnol* **35**: 997–1013.

Ladoukakis ED, Eyre-Walker A. 2008. The excess of small inverted repeats in prokaryotes. *J Mol Evol* **67**: 291–300.

Lafer EM, Moller A, Nordheim A, Stollar BD, Rich A. 1981. Antibodies specific for left-handed Z-DNA. *Proceedings of the National Academy of Sciences* **78**: 3546–3550.

Lagravère C, Malfoy B, Leng M, Laval J. 1984. Ring-opened alkylated guanine is not repaired in Z-DNA. *Nature* **310**: 798–800.

Lai PJ, Lim CT, Le HP, Katayama T, Leach DRF, Furukohri A, Maki H. 2016. Long inverted repeat transiently stalls DNA replication by forming hairpin structures on both leading and lagging strands. *Genes Cells* **21**: 136–145.

Lam EYN, Beraldi D, Tannahill D, Balasubramanian S. 2013. G-quadruplex structures are stable and detectable in human genomic DNA. *Nat Commun* **4**: 1796.

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**: 495–501.

Leach DRF. 1994. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* **16**: 893–900.

Leach DR, Okely EA, Pinder DJ. 1997. Repair by recombination of DNA containing a palindromic sequence. *Mol Microbiol* **26**: 597–606.

Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, et al. 2017. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**: 409–413.

Le DT, Uram JN, Wang H, Bartlett B, Kemberling H, Eyring A, Skora A, Azad NS, Laheru DA, Donehower RC, et al. 2015. PD-1 blockade in tumors with mismatch repair deficiency. *J Clin Oncol* **33**: LBA100–LBA100.

Lee JC-I, Tseng B, Ho B-C, Linacre A. 2015. pSTR Finder: a rapid method to discover polymorphic short tandem repeat markers from whole-genome sequences. *Investig Genet* **6**. http://dx.doi.org/10.1186/s13323-015-0027-x.

Lee TI, Young RA. 2013. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**: 1237–1251.

Leung M-Y, Choi KP, Xia A, Chen LHY. 2005. Nonrandom clusters of palindromes in herpesvirus genomes. *J Comput Biol* **12**: 331–354.

Levo M, Segal E. 2014. In pursuit of design principles of regulatory sequences. *Nat Rev Genet* **15**: 453–468.

Lieblein AL, Fürtig B, Schwalbe H. 2013. Optimizing the kinetics and thermodynamics of DNA i-motif folding. *Chembiochem* **14**: 1226–1230.

Li H, Xiao J, Li J, Lu L, Feng S, Dröge P. 2009. Human genomic Z-DNA segments probed by the Zα domain of ADAR1. *Nucleic Acids Res* **37**: 2737–2746.

Lilley DM. 1980. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc Natl Acad Sci U S A* **77**: 6468–6472.

Lillo F, Basile S, Mantegna RN. 2002. Comparative genomics study of inverted repeats in bacteria. *Bioinformatics* **18**: 971–979.

Lima WF, Murray HM, Damle SS, Hart CE, Hung G, De Hoyos CL, Liang X-H, Crooke ST. 2016. Viable RNaseH1 knockout mice show RNaseH1 is essential for R loop processing, mitochondrial and liver function. *Nucleic Acids Res* **44**: 5299–5312.

Lim B, Mun J, Kim YS, Kim S-Y. 2017. Variability in Chromatin Architecture and Associated DNA Repair at Genomic Positions Containing Somatic Mutations. *Cancer Res* **77**: 2822–2833.

Lim KW, Jenjaroenpun P, Low ZJ, Khong ZJ, Ng YS, Kuznetsov VA, Phan AT. 2015. Duplex stem-loop-containing quadruplex motifs in the human genome: a combined genomic and structural study. *Nucleic Acids Res* **43**: 5630–5646.

Lim KW, Khong ZJ, Phan AT. 2013. Thermal Stability of DNA Quadruplex–Duplex Hybrids. *Biochemistry* **53**: 247–257.

Lim KW, Lacroix L, Yue DJE, Lim JKC, Lim JMW, Phan AT. 2010. Coexistence of two distinct G-quadruplex conformations in the hTERT promoter. *J Am Chem Soc* **132**: 12331–12342.

Lim KW, Phan AT. 2013. Structural basis of DNA quadruplex-duplex junction formation. *Angew Chem Int Ed Engl* **52**: 8566–8569.

Lin S, Kowalski D. 1994. DNA helical instability facilitates initiation at the SV40 replication origin. *J Mol Biol* **235**: 496–507.

Lipps HJ, Nordheim A, Lafer EM, Ammermann D, David Stollar B, Rich A. 1983. Antibodies against Z DNA react with the macronucleus but not the micronucleus of the hypotrichous ciliate stylonychia mytilus. *Cell* **32**: 435–441.

Liu R, Liu H, Chen X, Kirby M, Brown PO, Zhao K. 2001. Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell* **106**: 309–318.

Li X, Heyer W-D. 2008. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res* **18**: 99–113.

Li X, Lindahl L, Sha Y, Zengel JM. 1997. Analysis of the Bacillus subtilis S10 ribosomal protein gene cluster identifies two promoters that may be responsible for transcription of the entire 15-kilobase S10-spc-alpha cluster. *J Bacteriol* **179**: 7046–7054.

Li Y, Syed J, Sugiyama H. 2016. RNA-DNA Triplex Formation by Long Noncoding RNAs. *Cell Chem Biol* **23**: 1325–1333.

Lobachev KS, Rattray A, Narayanan V. 2007. Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells. *Front Biosci* **12**: 4208–4220.

Lobachev KS, Shor BM, Tran HT, Taylor W, Keen JD, Resnick MA, Gordenin DA. 1998. Factors affecting inverted repeat stimulation of recombination and deletion in Saccharomyces cerevisiae. *Genetics* **148**: 1507–1524.

Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA. 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. *EMBO J* **19**: 3822–3830.

Lu L, Jia H, Dröge P, Li J. 2007. The human genome-wide distribution of DNA palindromes. *Funct Integr Genomics* **7**: 221–227.

Lu S, Wang G, Bacolla A, Zhao J, Spitser S, Vasquez KM. 2015. Short Inverted Repeats Are Hotspots for Genetic Instability: Relevance to Cancer Genomes. *Cell Rep*. http://dx.doi.org/10.1016/j.celrep.2015.02.039.

Lyamichev VI, Mirkin SM, Frank-Kamenetskii MD. 1985. A pH-dependent structural transition in the homopurine-homopyrimidine tract in superhelical DNA. *J Biomol Struct Dyn* **3**: 327–338.

Lyamichev VI, Mirkin SM, Frank-Kamenetskii MD. 1986. Structures of homopurine-homopyrimidine tract in superhelical DNA. *J Biomol Struct Dyn* **3**: 667–669.

Lyamichev VI, Panyutin IG, Frank-Kamenetskii MD. 1983. Evidence of cruciform structures in superhelical DNA provided by two-dimensional gel electrophoresis. *FEBS Lett* **153**: 298–302.

Ma J, Wang MD. 2016. DNA supercoiling during transcription. *Biophys Rev* **8**: 75–87.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

Marcel V, Tran PLT, Sagne C, Martel-Planche G, Vaslin L, Teulade-Fichou M-P, Hall J, Mergny J-L, Hainaut P, Van Dyck E. 2011. G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. *Carcinogenesis* **32**: 271–278.

Martadinata H, Phan AT. 2014. Formation of a stacked dimeric G-quadruplex containing bulges by the 5'-terminal region of human telomerase RNA (hTERC). *Biochemistry* **53**: 1595–1600.

Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ. 2014. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol* **15**: 465–481.

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**: 1029–1041.e21.

Mathelier A, Xin B, Chiu T-P, Yang L, Rohs R, Wasserman WW. 2016. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst* **3**: 278–286.e4.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**: 1190–1195.

McAlinden A, Havlioglu N, Liang L, Davies SR, Sandell LJ. 2005. Alternative Splicing of Type II Procollagen Exon 2 Is Regulated by the Combination of a Weak 5′ Splice Site and an Adjacent Intronic Stem-loop Cis Element. *J Biol Chem* **280**: 32700–32711.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**: 271–277.

Mergny JL, Lacroix L, Teulade-Fichou MP, Hounsou C, Guittat L, Hoarau M, Arimondo PB, Vigneron JP, Lehn JM, Riou JF, et al. 2001. Telomerase inhibitors based on quadruplex ligands selected by a fluorescence assay. *Proc Natl Acad Sci U S A* **98**: 3062–3067.

Mergny JL, Phan AT, Lacroix L. 1998. Following G-quartet formation by UV-spectroscopy. *FEBS Lett* **435**: 74–78.

Messer PW, Arndt PF. 2007. The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol* **24**: 1190–1197.

Mikheikin AL, Lushnikov AY, Lyubchenko YL. 2006. Effect of DNA supercoiling on the geometry of holliday junctions. *Biochemistry* **45**: 12998–13006.

Mirkin SM. 2007. Expandable DNA repeats and human disease. *Nature* **447**: 932–940.

Mirkin SM, Lyamichev VI, Drushlyak KN, Dobrynin VN, Filippov SA, Frank-Kamenetskii MD. 1987. DNA H form requires a homopurine–homopyrimidine mirror repeat. *Nature* **330**: 495–497.

Mitas M. 1997. Trinucleotide repeats associated with human disease. *Nucleic Acids Res* **25**: 2245–2253.

Mizuuchi K, Kemper B, Hays J, Weisberg RA. 1982. T4 endonuclease VII cleaves holliday structures. *Cell* **29**: 357–365.

Möller A, Gabriels JE, Lafer EM, Nordheim A, Rich A, Stollar BD. 1982. Monoclonal antibodies recognize different parts of Z-DNA. *J Biol Chem* **257**: 12081–12085.

Mondal T, Subhash S, Vaid R, Enroth S, Uday S, Reinius B, Mitra S, Mohammed A, James AR, Hoberg E, et al. 2015. MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* **6**: 7743.

Morganella S, Alexandrov LB, Glodzik D, Zou X, Davies H, Staaf J, Sieuwerts AM, Brinkman AB, Martin S, Ramakrishna M, et al. 2016. The topography of mutational processes in breast cancer genomes. *Nat Commun* **7**: 11383.

Morgan RK, Batra H, Gaerig VC, Hockings J, Brooks TA. 2016. Identification and characterization of a new G-quadruplex forming region within the kRAS promoter as a transcriptional regulator. *Biochim Biophys Acta* **1859**: 235–245.

Mortimer SA, Kidwell MA, Doudna JA. 2014. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**: 469–479.

Moser HE, Dervan PB. 1987. Sequence-specific cleavage of double helical DNA by triple helix formation. *Science* **238**: 645–650.

Mukundan VT, Phan AT. 2013. Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J Am Chem Soc* **135**: 5017–5028.

Murphy KM, Zhang S, Geiger T, Hafez MJ, Bacher J, Berg KD, Eshleman JR. 2006. Comparison of the Microsatellite Instability Analysis System and the Bethesda Panel for the Determination of Microsatellite Instability in Colorectal Cancers. *J Mol Diagn* **8**: 305–311.

Nag DK, Kurst A. 1997. A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast Saccharomyces cerevisiae. *Genetics* **146**: 835–847.

Nag DK, Petes TD. 1991. Seven-base-pair inverted repeats in DNA form stable hairpins in vivo in Saccharomyces cerevisiae. *Genetics* **129**: 669–673.

Nakanishi K, Shima A, Fukuda M, Fujita S. 1979. Age associated increase of single-stranded regions in the DNA of mouse brain and liver cells. *Mech Ageing Dev* **10**: 273–281.

Nambiar M, Goldsmith G, Moorthy BT, Lieber MR, Joshi MV, Choudhary B, Hosur RV, Raghavan SC. 2010. Formation of a G-quadruplex at the BCL2 major breakpoint region of the t(14;18) translocation in follicular lymphoma. *Nucleic Acids Res* **39**: 936–948.

Naughton C, Avlonitis N, Corless S, Prendergast JG, Mati IK, Eijk PP, Cockroft SL, Bradley M, Ylstra B, Gilbert N. 2013. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol* **20**: 387–395.

Neidle S. 2010. Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer. *FEBS J* **277**: 1118–1125.

Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, et al. 2012a. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**: 979–993.

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, et al. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**: 47–54.

Nik-Zainal S, Kucab JE, Morganella S, Glodzik D, Alexandrov LB, Arlt VM, Weninger A, Hollstein M, Stratton MR, Phillips DH. 2015. The genome as a record of environmental exposure. *Mutagenesis* **30**: 763–770.

Nik-Zainal S, Morganella S. 2017. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. *Clin Cancer Res* **23**: 2617–2629.

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, et al. 2012b. The life history of 21 breast cancers. *Cell* **149**: 994–1007.

Nordheim A, Lafer EM, Peck LJ, Wang JC, Stollar BD, Rich A. 1982. Negatively supercoiled plasmids contain left-handed Z-DNA segments as detected by specific antibody binding. *Cell* **31**: 309–318.

Nordheim A, Rich A. 1983. Negatively supercoiled simian virus 40 DNA contains Z-DNA segments within transcriptional enhancer sequences. *Nature* **303**: 674–679.

O'Dushlaine CT, Shields DC. 2008. Marked variation in predicted and observed variability of tandem repeat loci across the human genome. *BMC Genomics* **9**: 175.

Oganesian L, Bryan TM. 2007. Physiological relevance of telomeric G-quadruplex formation: a potential drug target. *Bioessays* **29**: 155–165.

Oganesian L, Moon IK, Bryan TM, Jarstfer MB. 2006. Extension of G-quadruplex DNA by ciliate telomerase. *EMBO J* **25**: 1148–1159.

Ohno M, Fukagawa T, Lee JS, Ikemura T. 2002. Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma* **111**: 201–213.

Onel B, Carver M, Wu G, Timonina D, Kalarn S, Larriva M, Yang D. 2016. A New G-Quadruplex with Hairpin Loop Immediately Upstream of the Human BCL2 P1 Promoter Modulates Transcription. *J Am Chem Soc* **138**: 2563–2570.

Palumbo SL, Ebbinghaus SW, Hurley LH. 2009. Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. *J Am Chem Soc* **131**: 10878–10891.

Panayotatos N, Wells RD. 1981. Cruciform structures in supercoiled DNA. *Nature* **289**: 466–470.

Panigrahi GB, Lau R, Montgomery SE, Leonard MR, Pearson CE. 2005. Slipped (CTG)*(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. *Nat Struct Mol Biol* **12**: 654–662.

Pearson CE, Eichler EE, Lorenzetti D, Kramer SF, Zoghbi HY, Nelson DL, Sinden RR. 1998. Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry* **37**: 2701–2708.

Pearson CE, Sinden RR. 1996. Alternative Structures in Duplex DNA Formed within the Trinucleotide Repeats of the Myotonic Dystrophy and Fragile X Loci†. *Biochemistry* **35**: 5041–5053.

Pearson CE, Zorbas H, Price GB, Zannis-Hadjopoulos M. 1996. Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J Cell Biochem* **63**: 1–22.

Peck LJ, Nordheim A, Rich A, Wang JC. 1982. Flipping of cloned d(pCpG)n.d(pCpG)n DNA sequences from right- to left-handed helical structure by salt, Co(III), or negative supercoiling. *Proceedings of the National Academy of Sciences* **79**: 4560–4564.

Peck LJ, Wang JC. 1983. Energetics of B-to-Z transition in DNA. *Proc Natl Acad Sci U S A* **80**: 6206–6210.

Pfeifer GP, You Y-H, Besaratinia A. 2005. Mutations induced by ultraviolet light. *Mutat Res/Fundam Mol Mech Mutag* **571**: 19–31.

Phan AT, Kuryavyi V, Burge S, Neidle S, Patel DJ. 2007. Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J Am Chem Soc* **129**: 4386–4392.

Piazza A, Adrian M, Samazan F, Heddi B, Hamon F, Serero A, Lopes J, Teulade-Fichou M-P, Phan AT, Nicolas A. 2015. Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *EMBO J* **34**: 1718–1734.

Pickrell JK, Gaffney DJ, Gilad Y, Pritchard JK. 2011. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**: 2144–2146.

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, et al. 2010a. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**: 191–196.

Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C, et al. 2010b. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**: 184–190.

Pokrzywa R, Polanski A. 2010. BWtrs: A tool for searching for tandem repeats in DNA sequences based on the Burrows-Wheeler transform. *Genomics* **96**: 316–321.

Polak P, Karlić R, Koren A, Thurman R, Sandstrom R, Lawrence M, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA, et al. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**: 360–364.

Poon SL, Huang MN, Choo Y, McPherson JR, Yu W, Heng HL, Gan A, Myint SS, Siew EY, Ler LD, et al. 2015. Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med* **7**: 38.

Postepska-Igielska A, Giwojna A, Gasri-Plotnitsky L, Schmitt N, Dold A, Ginsberg D, Grummt I. 2015. LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. *Mol Cell* **60**: 626–636.

Prakash R, Zhang Y, Feng W, Jasin M. 2015. Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb Perspect Biol* **7**: a016600.

Press MO, Carlson KD, Queitsch C. 2014. The overdue promise of short tandem repeat variation for heritability. *Trends Genet* **30**: 504–512.

Price GB, Modak SP, Makinodan T. 1971. Age-Associated Changes in the DNA of Mouse Tissue. *Science* **171**: 917–920.

Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ. 2016. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res* **44**: 3750–3762.

Rachwal PA, Stuart Findlow I, Werner JM, Brown T, Fox KR. 2007. Intramolecular DNA quadruplexes with different arrangements of short and long loops. *Nucleic Acids Res* **35**: 4214–4222.

Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh C-L, Haworth IS, Lieber MR. 2005. Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. *J Biol Chem* **280**: 22749–22760.

Raghavan SC, Tsai A, Hsieh C, Lieber MR. 2006. Analysis of Non-B DNA Structure at Chromosomal Sites in the Mammalian Genome. In *Methods in Enzymology*, pp. 301–316.

Rahmouni AR, Wells RD. 1989. Stabilization of Z DNA in vivo by localized supercoiling. *Science* **246**: 358–363.

Raiber E-A, Kranaster R, Lam E, Nikan M, Balasubramanian S. 2012. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res* **40**: 1499–1508.

Rankin S, Reszka AP, Huppert J, Zloh M, Parkinson GN, Todd AK, Ladame S, Balasubramanian S, Neidle S. 2005. Putative DNA quadruplex formation within the human c-kit oncogene. *J Am Chem Soc* **127**: 10584–10589.

Ray BK, Dhar S, Henry C, Rich A, Ray A. 2013. Epigenetic regulation by Z-DNA silencer function controls cancer-associated ADAM-12 expression in breast cancer: cross-talk between MeCP2 and NF1 transcription factor family. *Cancer Res* **73**: 736–744.

Ray BK, Dhar S, Shakya A, Ray A. 2011. Z-DNA-forming silencer in the first exon regulates human ADAM-12 gene expression. *Proc Natl Acad Sci U S A* **108**: 103–108.

Reilly SM, Morgan RK, Brooks TA, Wadkins RM. 2015. Effect of interior loop length on the thermal stability and pK(a) of i-motif DNA. *Biochemistry* **54**: 1364–1370.

Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, van Swieten JC, Myllykangas L, et al. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**: 257–268.

Rentzeperis D, Alessi K, Marky LA. 1993. Thermodynamics of DNA hairpins: contribution of loop size to hairpin stability and ethidium binding. *Nucleic Acids Res* **21**: 2683–2689.

Rhodes D, Lipps HJ. 2015. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**: 8627–8637.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Rich A, Zhang S. 2003. Z-DNA: the long road to biological function. *Nat Rev Genet* **4**: 566–572.

Richman S. 2015. Deficient mismatch repair: Read all about it (Review). *Int J Oncol* **47**: 1189–1202.

Riggi N, Knoechel B, Gillespie SM, Rheinbay E, Boulay G, Suvà ML, Rossetti NE, Boonseng WE, Oksuz O, Cook EB, et al. 2014. EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in Ewing sarcoma. *Cancer Cell* **26**: 668–681.

Rimokh R, Rouault JP, Wahbi K, Gadoux M, Lafage M, Archimbaud E, Charrin C, Gentilhomme O, Germain D, Samarut J. 1991. A chromosome 12 coding region is juxtaposed to the MYC protooncogene locus in a t(8;12)(q24;q22) translocation in a case of B-cell chronic lymphocytic leukemia. *Genes Chromosomes Cancer* **3**: 24–36.

Risitano A, Fox KR. 2003. Stability of intramolecular DNA quadruplexes: comparison with DNA duplexes. *Biochemistry* **42**: 6507–6513.

Roberts R, Crothers D. 1992. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science* **258**: 1463–1466.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**: 970–976.

Rocca R, Talarico C, Moraca F, Costa G, Romeo I, Ortuso F, Alcaro S, Artese A. 2017. Molecular recognition of a carboxy pyridostatin toward G-quadruplex structures: Why does it prefer RNA? *Chem Biol Drug Des* **90**: 919–925.

Rodriguez R, Müller S, Yeoman JA, Trentesaux C, Riou J-F, Balasubramanian S. 2008. A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. *J Am Chem Soc* **130**: 15758–15759.

Rogers FA, Lloyd JA, Glazer PM. 2005. Triplex-forming oligonucleotides as potential tools for modulation of gene expression. *Curr Med Chem Anticancer Agents* **5**: 319–326.

Rooney SM, Moore PD. 1995. Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proc Natl Acad Sci U S A* **92**: 2141–2144.

Rosche WA, Trinh TQ, Sinden RR. 1997. Leading strand specific spontaneous mutation corrects a quasipalindrome by an intermolecular strand switch mechanism. *J Mol Biol* **269**: 176–187.

Rothenburg S, Koch-Nolte F, Haag F. 2001. DNA methylation and Z-DNA formation as mediators of quantitative differences in the expression of alleles. *Immunol Rev* **184**: 286–298.

Rougée M, Faucon B, Mergny JL, Barcelo F, Giovannangeli C, Garestier T, Hélène C. 1992. Kinetics and thermodynamics of triple-helix formation: effects of ionic strength and mismatches. *Biochemistry* **31**: 9269–9278.

Rouleau S, Glouzon J-PS, Brumwell A, Bisaillon M, Perreault J-P. 2017. 3’ UTR G-quadruplexes regulate miRNA binding. *RNA* **23**: 1172–1179.

Rusling DA, Broughton-Head VJ, Tuck A, Khairallah H, Osborne SD, Brown T, Fox KR. 2008. Kinetic studies on the formation of DNA triplexes containing the nucleoside analogue 2′-O-(2-aminoethyl)-5-(3-amino-1-propynyl)uridine. *Org Biomol Chem* **6**: 122–129.

Rusling DA, Peng G, Srinivasan N, Fox KR, Brown T. 2009. DNA triplex formation with 5-dimethylaminopropargyl deoxyuridine. *Nucleic Acids Res* **37**: 1288–1296.

Sabarinathan R, Mularoni L, Deu-Pons J, Gonzalez-Perez A, Lopez-Bigas N. 2015. Nucleotide excision repair is impaired by binding of transcription factors to DNA. http://dx.doi.org/10.1101/028886.

Saglio G, Grazia Borrello M, Guerrasio A, Sozzi G, Serra A, di Celle PF, Foa R, Ferrarini M, Roncella S, Borgna Pignatti C. 1993. Preferential clustering of chromosomal breakpoints in Burkitt's lymphomas and L3 type acute lymphoblastic leukemias with a t(8;14) translocation. *Genes Chromosomes Cancer* **8**: 1–7.

Samadashwily GM, Raca G, Mirkin SM. 1997. Trinucleotide repeats affect DNA replication in vivo. *Nat Genet* **17**: 298–304.

Sankar TS, Sabari Sankar T, Wastuwidyaningtyas BD, Dong Y, Lewis SA, Wang JD. 2016. The nature of mutations induced by replication–transcription collisions. *Nature* **535**: 178–181.

Santos-Pereira JM, Aguilera A. 2015. R loops: new modulators of genome dynamics and function. *Nat Rev Genet* **16**: 583–597.

Sarai A, Sugiura S, Torigoe H, Shindo H. 1993. Thermodynamic and kinetic analyses of DNA triplex formation: application of filter-binding assay. *J Biomol Struct Dyn* **11**: 245–252.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One* **8**: e54710.

Sawaya SM, Bagshaw AT, Buschiazzo E, Gemmell NJ. 2012. Promoter Microsatellites as Modulators of Human Gene Expression. In *Advances in Experimental Medicine and Biology*, pp. 41–54.

Schroth GP, Chou PJ, Ho PS. 1992. Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J Biol Chem* **267**: 11846–11855.

Schroth GP, Ho PS. 1995. Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res* **23**: 1977–1983.

Schuster-Böckler B, Lehner B. 2012. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**: 504–507.

Schwartz T, Rould MA, Lowenhaupt K, Herbert A, Rich A. 1999. Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science* **284**: 1841–1845.

Sen D, Gilbert W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334**: 364–366.

Sen D, Gilbert W. 1990. A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature* **344**: 410–414.

Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y. 1999. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett* **455**: 70–74.

Shin S-I, Ham S, Park J, Seo SH, Lim CH, Jeon H, Huh J, Roh T-Y. 2016. Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. *DNA Res*. http://dx.doi.org/10.1093/dnares/dsw031.

Shrestha P, Xiao S, Dhakal S, Tan Z, Mao H. 2014. Nascent RNA transcripts facilitate the formation of G-quadruplexes. *Nucleic Acids Res* **42**: 7236–7246.

Siddiqui-Jain A, Grand CL, Bearss DJ, Hurley LH. 2002. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* **99**: 11593–11598.

Sinden RR, Pettijohn DE. 1984. Cruciform transitions in DNA. *J Biol Chem* **259**: 6593–6600.

Sinden RR, Richard, Sinden R. 2007. Slipped strand DNA structures. *Front Biosci* **12**: 4788.

Sinden RR, Wells RD. 1992. DNA structure, mutations, and human genetic disease. *Curr Opin Biotechnol* **3**: 612–622.

Sinden RR, Zheng GX, Brankamp RG, Allen KN. 1991. On the deletion of inverted repeated DNA in Escherichia coli: effects of length, thermal stability, and cruciform formation in vivo. *Genetics* **129**: 991–1005.

Singleton CK, Klysik J, Stirdivant SM, Wells RD. 1982. Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. *Nature* **299**: 312–316.

Soumpasis DM, Robert-Nicoud M, Jovin TM. 1987. B-Z DNA conformational transition in 1:1 electrolytes: dependence upon counterion size. *FEBS Lett* **213**: 341–344.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. *Nature* **458**: 719–724.

Strawbridge EM, Benson G, Gelfand Y, Benham CJ. 2010. The distribution of inverted repeat sequences in the Saccharomyces cerevisiae genome. *Curr Genet* **56**: 321–340.

Sun D, Hurley LH. 2009. The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *J Med Chem* **52**: 2863–2874.

Sun H, Karow JK, Hickson ID, Maizels N. 1998. The Bloom's Syndrome Helicase Unwinds G4 DNA. *J Biol Chem* **273**: 27587–27592.

Takahama K, Kino K, Arai S, Kurokawa R, Oyoshi T. 2011. Identification of Ewing's sarcoma protein as a G-quadruplex DNA- and RNA-binding protein. *FEBS J* **278**: 988–998.

Takahashi K, Yamanaka S. 2006. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**: 663–676.

Takahashi S, Sugimoto N. 2015. Pressure-dependent formation of i-motif and G-quadruplex DNA structures. *Phys Chem Chem Phys* **17**: 31004–31010.

Tam M, Erin Montgomery S, Kekis M, Stollar BD, Price GB, Pearson CE. 2003. Slipped (CTG).(CAG) repeats of the myotonic dystrophy locus: surface probing with anti-DNA antibodies. *J Mol Biol* **332**: 585–600.

Tanay A, Siggia ED. 2008. Sequence context affects the rate of short insertions and deletions in flies and primates. *Genome Biol* **9**: R37.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Thomas TJ, Gunnia UB, Thomas T. 1991. Polyamine-induced B-DNA to Z-DNA conformational transition of a plasmid DNA with (dG-dC)n insert. *J Biol Chem* **266**: 6137–6141.

Tian T, Chen Y-Q, Wang S-R, Zhou X. 2018. G-Quadruplex: A Regulator of Gene Expression and Its Chemical Targeting. *Chem* **4**: 1314–1344.

Tippana R, Xiao W, Myong S. 2014. G-Quadruplex Folding Depends on its Loop Size and Sequence: Extreme Fast Folding Kinetics Observed in Human Telomere and its Isomer. *Biophys J* **106**: 64a.

Tomkova M, Tomek J, Kriaucionis S, Schuster-Böckler B. 2018. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol* **19**: 129.

Tran H, Degtyareva N, Gordenin D, Resnick MA. 1997. Altered replication and inverted repeats induce mismatch repair-independent recombination between highly diverged DNAs in yeast. *Mol Cell Biol* **17**: 1027–1036.

van Noort V, Worning P, Ussery DW, Rosche WA, Sinden RR. 2003. Strand misalignments lead to quasipalindrome correction. *Trends Genet* **19**: 365–369.

Varani G. 1995. Exceptionally stable nucleic acid hairpins. *Annu Rev Biophys Biomol Struct* **24**: 379–404.

Varizhuk A, Ischenko D, Tsvetkov V, Novikov R, Kulemin N, Kaluzhny D, Vlasenok M, Naumov V, Smirnov I, Pozmogova G. 2017. The expanding repertoire of G4 DNA structures. *Biochimie* **135**: 54–62.

Vasquez KM, Narayanan L, Glazer PM. 2000. Specific mutations induced by triplex-forming oligonucleotides in mice. *Science* **290**: 530–533.

Vasquez KM, Wilson JH. 1998. Triplex-directed modification of genes and gene activity. *Trends Biochem Sci* **23**: 4–9.

Vilar E, Gruber SB. 2010. Microsatellite instability in colorectal cancer-the stable evidence. *Nat Rev Clin Oncol* **7**: 153–162.

Vilenchik MM, Knudson AG Jr. 2000. Inverse radiation dose-rate effects on somatic and germ-line mutations and DNA damage rates. *Proc Natl Acad Sci U S A* **97**: 5381–5386.

Vogelstein B, Lane D, Levine AJ. 2000. Surfing the p53 network. *Nature* **408**: 307–310.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer Genome Landscapes. *Science* **339**: 1546–1558.

Voineagu I, Narayanan V, Lobachev KS, Mirkin SM. 2008. Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc Natl Acad Sci U S A* **105**: 9936–9941.

Voloshin ON, Mirkin SM, Lyamichev VI, Belotserkovskii BP, Frank-Kamenetskii MD. 1988. Chemical probing of homopurine-homopyrimidine mirror repeats in supercoiled DNA. *Nature* **333**: 475–476.

Vorlíčková M, Kejnovská I, Sagi J, Renčiuk D, Bednářová K, Motlová J, Kypr J. 2012. Circular dichroism and guanine quadruplexes. *Methods* **57**: 64–75.

Wallace SS. 2014. Base excision repair: a critical player in many games. *DNA Repair* **19**: 14–26.

Wang AH, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, Rich A. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282**: 680–686.

Wang G, Carbajal S, Vijg J, DiGiovanni J, Vasquez KM. 2008. DNA structure-induced genomic instability in vivo. *J Natl Cancer Inst* **100**: 1815–1817.

Wang G, Christensen LA, Vasquez KM. 2006. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci U S A* **103**: 2677–2682.

Wang G, Seidman MM, Glazer PM. 1996. Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. *Science* **271**: 802–805.

Wang G, Vasquez K. 2017. Effects of Replication and Transcription on DNA Structure-Related Genetic Instability. *Genes* **8**: 17.

Wang G, Vasquez KM. 2009. Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. *Mol Carcinog* **48**: 286–298.

Wang G, Vasquez KM. 2004. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc Natl Acad Sci U S A* **101**: 13448–13453.

Wang G, Vasquez KM. 2006. Non-B DNA structure-induced genetic instability. *Mutat Res* **598**: 103–119.

Wang Y, Huang J-M. 2017. Lirex: A Package for Identification of Long Inverted Repeats in Genomes. *Genomics Proteomics Bioinformatics* **15**: 141–146.

Wang Y, Patel DJ. 1993. Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure* **1**: 263–282.

Wanrooij PH, Uhler JP, Shi Y, Westerlund F, Falkenberg M, Gustafsson CM. 2012. A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary stability of the mitochondrial R-loop. *Nucleic Acids Res* **40**: 10334–10344.

Wanrooij PH, Uhler JP, Simonsson T, Falkenberg M, Gustafsson CM. 2010. G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. *Proc Natl Acad Sci U S A* **107**: 16072–16077.

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* **14**: 1861–1869.

Watson JD, Crick FHC. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**: 737–738.

Watson J, Hays FA, Ho PS. 2004. Definitions and analysis of DNA Holliday junction geometry. *Nucleic Acids Res* **32**: 3017–3027.

Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. 2014. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**: 1160–1165.

Wells RD. 2008. DNA triplexes and Friedreich ataxia. *FASEB J* **22**: 1625–1634.

Wells RD, Collier DA, Hanvey JC, Shimizu M, Wohlrab F. 1988. The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. *FASEB J* **2**: 2939–2949.

Wilda M, Busch K, Klose I, Keller T, Woessmann W, Kreuder J, Harbott J, Borkhardt A. 2004. Level of MYC overexpression in pediatric Burkitt's lymphoma is strongly dependent on genomic breakpoint location within the MYC locus. *Genes Chromosomes Cancer* **41**: 178–182.

Wittig B, Dorbic T, Rich A. 1991. Transcription is associated with Z-DNA formation in metabolically active permeabilized mammalian cell nuclei. *Proc Natl Acad Sci U S A* **88**: 2259–2263.

Wittig B, Wölfl S, Dorbic T, Vahrson W, Rich A. 1992. Transcription of human c-myc in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene. *EMBO J* **11**: 4653–4663.

Wojcik EA, Brzostek A, Bacolla A, Mackiewicz P, Vasquez KM, Korycka-Machala M, Jaworski A, Dziadek J. 2012. Direct and inverted repeats elicit genetic instability by both exploiting and eluding DNA double-strand break repair systems in mycobacteria. *PLoS One* **7**: e51064.

Wölfl S, Martinez C, Rich A, Majzoub JA. 1996. Transcription of the human corticotropin-releasing hormone gene in NPLC cells is correlated with Z-DNA formation. *Proc Natl Acad Sci U S A* **93**: 3664–3668.

Wölfl S, Wittig B, Rich A. 1995. Identification of transcriptionally induced Z-DNA segments in the human c-myc gene. *Biochim Biophys Acta* **1264**: 294–302.

Wong B, Chen S, Kwon J-A, Rich A. 2007. Characterization of Z-DNA as a nucleosome-boundary element in yeast Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **104**: 2229–2234.

Wong HM, Stegle O, Rodgers S, Huppert JL. 2010. A toolbox for predicting g-quadruplex formation and stability. *J Nucleic Acids* **2010**. http://dx.doi.org/10.4061/2010/564946.

Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318**: 1108–1113.

Woodside MT, Behnke-Parks WM, Larizadeh K, Travers K, Herschlag D, Block SM. 2006. Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins. *Proc Natl Acad Sci U S A* **103**: 6190–6195.

Wu Y, Rawtani N, Thazhathveetil AK, Kenny MK, Seidman MM, Brosh RM. 2008. Human Replication Protein A Melts a DNA Triple Helix Structure in a Potent and Specific Manner†. *Biochemistry* **47**: 5068–5077.

Wyman AR, White R. 1980. A highly polymorphic locus in human DNA. *Proc Natl Acad Sci U S A* **77**: 6754–6758.

Yáñez-Cuna JO, Arnold CD, Stampfel G, Boryń LM, Gerlach D, Rath M, Stark A. 2014. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res* **24**: 1147–1156.

Ye C, Ji G, Li L, Liang C. 2014. detectIR: a novel program for detecting perfect and imperfect inverted repeats using complex numbers and vector calculation. *PLoS One* **9**: e113349.

Zacharias W, Jaworski A, Wells RD. 1990. Cytosine methylation enhances Z-DNA formation in vivo. *J Bacteriol* **172**: 3278–3283.

Zahler AM, Williamson JR, Cech TR, Prescott DM. 1991. Inhibition of telomerase by G-quartet DNA structures. *Nature* **350**: 718–720.

Zahran M, Sevim Bayrak C, Elmetwaly S, Schlick T. 2015. RAG-3D: a search tool for RNA 3D substructures. *Nucleic Acids Res* **43**: 9474–9488.

Zannis-Hadjopoulos M, Frappier L, Khoury M, Price GB. 1988. Effect of anti-cruciform DNA monoclonal antibodies on DNA replication. *EMBO J* **7**: 1837–1844.

Zaytseva O, Quinn LM. 2018. DNA Conformation Regulates Gene Expression: The MYC Promoter and Beyond. *Bioessays* **40**: e1700235.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. 2015. The ensembl regulatory build. *Genome Biol* **16**: 56.

Zhang R, Lin Y, Zhang C-T. 2008. Greglist: a database listing potential G-quadruplex regulated genes. *Nucleic Acids Res* **36**: D372–6.

Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* **67**: 43–62.

Zhao J, Wang G, Del Mundo IM, McKinney JA, Lu X, Bacolla A, Boulware SB, Zhang C, Zhang H, Ren P, et al. 2018. Distinct Mechanisms of Nuclease-Directed DNA-Structure-Induced Genetic Instability in Cancer Genomes. *Cell Rep* **22**: 1200–1210.

Zheng GX, Kochel T, Hoepfner RW, Timmons SE, Sinden RR. 1991. Torsionally tuned cruciform and Z-DNA probes for measuring unrestrained supercoiling at specific sites in DNA of living cells. *J Mol Biol* **221**: 107–122.

Zheng GX, Sinden RR. 1988. Effect of base composition at the center of inverted repeated DNA sequences on cruciform transitions in DNA. *J Biol Chem* **263**: 5356–5361.

Zheng K-W, Xiao S, Liu J-Q, Zhang J-Y, Hao Y-H, Tan Z. 2013. Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control. *Nucleic Acids Res* **41**: 5533–5541.

Zhu Q, Hoong N, Aslanian A, Hara T, Benner C, Heinz S, Miga KH, Ke E, Verma S, Soroczynski J, et al. 2018. Heterochromatin-Encoded Satellite RNAs Induce Breast Cancer. *Mol Cell* **70**: 842–853.e7.

Zhu Q, Pao GM, Huynh AM, Suh H, Tonnu N, Nederlof PM, Gage FH, Verma IM. 2011. BRCA1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature* **477**: 179–184.

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**: 65–70.

Zou X, Morganella S, Glodzik D, Davies H, Li Y, Stratton MR, Nik-Zainal S. 2017. Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res* **45**: 11213–11221.